UNIVERSIDAD AUTONOMA DE MADRID

Escuela Politécnica Superior

Dpto. de Ingeniería Informática

Doctorado en Ingeniería Informática y Telecomunicación

Doctoral Thesis

Evolutionary Computation for Overlapping Community Detection in Social and Graph-based Information

Author
GEMA BELLO ORGAZ

Advisor
Dr. D. DAVID CAMACHO FERNÁNDEZ

June 2017

Department: Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid (UAM)
SPAIN

PhD Thesis: “Evolutionary Computation for Overlapping Community Detection in Social and Graph-based Information”

Author: **Gema Bello Orgaz**
Ingeniera en Informática
Universidad Carlos III de Madrid, Spain

Advisor: **David Camacho Fernández**
Doctor Ingeniero en Informática
(Universidad Autónoma de Madrid)
Universidad Autónoma de Madrid, SPAIN

Year: 2017

Committee: President:

Secretary:

Vocal 1:

Vocal 2:

Vocal 3:

A mi madre,

la persona más buena, luchadora, alegre

y valiente que he conocido. Mi ejemplo de vida.

Aunque ya no pueda estar a mi lado,

siempre está conmigo, dentro de mi corazón.

Abstract

The Community Detection Problem (CDP) in Social Networks has been widely studied from different areas such as Data Mining, Graph Mining or Social Network Analysis, amongst others. Nowadays, this problem has become highly relevant due to the growing interest in Social Networks, and its possible application in several disciplines such as sociology, biology, neuroscience, or computer science, whose information can be easily represented as networks or graphs.

This problem can be formulated as a graph-clustering problem, where the nodes belonging to a graph should be partitioned into groups according to the network topology. The partitioning of a graph is usually a division of the graph where each node belongs to only one community. However, a common feature observed in real-world networks is the existence of overlapping communities, where a particular node can belong to more than one community. When the Community Detection Algorithms (CDA) are applied to process social-based information, it is quite usual for a node of the network to belong to different groups, or communities. Therefore, it is quite interesting that this type of algorithms can deal with this feature, detecting communities of nodes with overlaps.

This dissertation has been focused on the design of an evolutionary approach, which can provide new methods to find overlapping and stable communities in a graph, improving the currently existing techniques in the state of the art. For this purpose, several encoding approaches have been designed, as well as several fitness functions that guide the searching process using some measures related to the graph theory. The different algorithms designed and implemented in this theses have been divided in two generations.

In order to combine the basic idea of classical clustering methods with the graph clustering methods, a first generation of algorithms have been designed. These algorithms adopt an evolutionary approach to optimize the search of communities, using different fitness functions which combine distances metrics based on features of the nodes, with measures related to the network topology of the graph (GCF). Analysing the experimental results obtained by these algorithms, it can be noticed that they present the same problem as the classical CDAs, related to their dependence on the graph structure. Therefore, to handle with this problem, these algorithms have been extended trying to promote solution diversity, in a new generation of algorithms.

The second generation of algorithms designed are based on Multi-Objective Genetic Algorithms (MOGA-OCD). The fitness function implemented in MOGA-OCD algorithm optimizes two classical objective functions in CDP; the first one is used to maximize the internal connectivity of the communities, whereas the second one is used to minimize the external connections to the rest of the graph. To select the most appropriate metrics for these objective functions, a comparative assessment of several connectivity metrics has been carried out using a real-world network. Finally, the algorithm has been evaluated against other well-known algorithms from the state of the art in CDP. The experimental results show that the proposed algorithms obtain good results for the different types of graphs with different structures. In addition, they improve overall the accuracy and quality of current approaches in CDP, showing its effectiveness as a new method for detecting overlapping communities.

Finally, and in order to evaluate the CDA algorithms designed in a real domain, this dissertation also includes the application of them to identify opinion communities related to Public Healthcare topics in Social Networks. 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 about 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. This thesis shows how CDAs can be applied to discover and track discussion communities on vaccination using information gathered from Twitter. In addition, this fact provides useful information that Public Healthcare Organizations may try to use for avoiding or mitigating new outbreaks of eradicated diseases.

Resumen

El Problema de Detección de Comunidades (o CDP del inglés Community Detection Problem) ha sido ampliamente estudiado desde diferentes áreas como la Minería de Datos, Minería de Grafos o Análisis de Redes Sociales, entre otros. Hoy en día, este problema se ha vuelto muy relevante debido al creciente interés en las redes sociales y su posible aplicación en varias disciplinas como la sociología, la biología, la neurociencia o la informática, cuya información puede representarse fácilmente como redes o grafos.

La idea principal de este problema puede considerarse similar a la del clustering de grafos, donde los nodos pertenecientes a un grafo deben ser divididos en grupos de acuerdo con la topología de red. La partición de un grafo generalmente consiste en una división del grafo donde cada nodo pertenece solamente a una comunidad. Sin embargo, una característica común observada en las redes del mundo real es la existencia de comunidades superpuestas, donde un nodo particular puede pertenecer a varias comunidades. Cuando los algoritmos de detección de comunidades se están aplicando para procesar información de base social, es muy habitual que un nodo de la red pertenezca a diferentes grupos o comunidades. Por lo tanto, es muy interesante que este tipo de algoritmos puedan manejar esta característica, detectando comunidades de nodos con superposiciones.

Esta tesis se centra en proponer un enfoque evolutivo que pueda aportar nuevos métodos para encontrar comunidades superpuestas y estables en un grafo, mejorando las técnicas actuales del estado del arte. Para este propósito, se han diseñado varias codificaciones diferentes, así como varias funciones de aptitud que guíen el proceso de búsqueda usando diferente medidas relacionadas con la teoría de los grafos. Los diferentes algoritmos diseñados e implementados en estas tesis se han dividido en dos generaciones.

Combinando la idea básica de los métodos clásicos de clustering, con los métodos de clustering para grafos, se ha desarrollado la primera generación de algoritmos. Estos algoritmos adoptan un enfoque evolutivo para optimizar la búsqueda de comunidades, utilizando diferentes funciones de aptitud, que combinan métricas de distancias basadas en características de los nodos, con medidas relacionadas con la topología del grafo (GCF). Analizando los resultados experimentales obtenidos por estos algoritmos, se puede observar que presentan el mismo problema que los algoritmos de detección de comunidades clásicos, en relación con su dependencia a la estructura del grafo. Por este motivo, con el objetivo de evitar este problema, dichos algoritmos se extienden tratando de promover una mayor diversidad de soluciones, y dando lugar a una segunda generación de algoritmos.

La segunda generación de algoritmos diseñados se basa en Algoritmos Genéticos Multi-Objetivos (MOGA-OCD). La función de aptitud implementada para estos algoritmos optimiza simultáneamente dos objetivos: El primero tratando de maximizar la conectividad interna de las comunidades, mientras que el segundo se utiliza para minimizar las conexiones externas con el resto de nodos del grafo. Para seleccionar las métricas más apropiadas para estas funciones objetivo, se ha realizado una evaluación comparativa de varias métricas de conectividad utilizando una red representado

un problema real. Y finalmente, el algoritmo ha sido evaluado frente a otros algoritmos conocidos del estado del arte, donde se ha podido comprobar que los nuevos algoritmos propuestos tienen buenos resultados para diferentes tipos de grafo, con estructuras diferentes. Además, estos resultados muestran que en general mejoran tanto la precisión, como la calidad respecto a los enfoques clásicos, mostrando así su efectividad como nuevos método para detectar comunidades solapadas.

Finalmente, con el fin de evaluar los algoritmos diseñados en un dominio real, esta tesis también incluye la aplicación de los mismos para identificar comunidades de opinión relacionadas con temas de Salud Pública en Redes Sociales. Las vacunas han contribuido a disminuir drásticamente la mortalidad por enfermedades infecciosas en el siglo XX. Sin embargo, en los últimos años varios movimientos sociales hablando en contra de las vacunas han surgido, influyendo en la opinión de la población sobre la vacunación. Estas comunidades que discuten acerca de las vacunas han aprovechado las redes sociales para difundir eficazmente sus teorías. Actualmente, brotes recientes de enfermedades ya casi erradicadas como el sarampión o la polio han demostrado el efecto negativo que tiene la disminución de las tasas de vacunación. Las Redes Sociales son una de las fuentes más importantes de Big Data, especialmente Twitter que genera más de 400 millones de tweets cada día. Por este motivo, esta tesis muestra cómo los algoritmos de detección de comunidades pueden ser aplicados para descubrir y rastrear estas comunidades de discusión sobre la vacunación con datos extraídos de Twitter. Además, este análisis muestra como estos algoritmos pueden proporcionar información útil, que las organizaciones de salud pública pueden tratar de utilizar para evitar o mitigar nuevos brotes de enfermedades.

Agradecimientos

Llegado a este momento, en el que el trabajo más duro parece que ya ha terminado, creo que es muy bonito poder mirar hacia atrás, y ver cuantas cosas has vivido, y con cuánta gente las has compartido. Echando un vistazo atrás, lo primero me sale es una sonrisa, recordando a toda esa gente que ha estado a mi lado durante estos últimos años, y a los que quiero dar las gracias.

En primer lugar, a mi familia que es el pilar fundamental de mi vida. Mi padre Juan, mis hermanos Carlos y Ricardo, y mis cuñadas Mary y Cristina. Por todas esas charlas en la cocina, risas, abrazos de armonía, consejos, y todo el apoyo y cariño que siempre me dais. Y por supuesto, a las cuatro personas más bonitas que existen en el mundo, mis sobrinos: Laura, Nuria, Elena y Victor. Le dais la magia y la alegría a mi vida, convirtiendo muchos momentos que compartimos en especiales, y simplemente felices. Sin todos vosotros no lo habría logrado.

A mi segunda familia, mis amigas: Esther, Gema, Marimar y Raquel. Sois las personas con las que comparto todo, las que siempre están ahí, para reír, llorar, hacer el payaso, desahogarme, salir de cañas, aconsejarme... Pensándolo bien, ¡no sé ni cómo me habéis podido aguantar tantos años!. Mil veces gracias por formar parte de mi vida.

Siguiendo con los amigos, a Silvia, que a pesar de estar geográficamente lejos, siempre la siento cercana a mí. Ella llego justo cuando decidí comenzar la aventura de la Tesis. Una simple casualidad nos unió, y desde entonces se ha convertido en una gran amiga. Que además últimamente ha tenido que sufrir muchos mis agobios infinitos con esto de intentar terminar y no ver el fin. También a mi amiga y compi de viajes Vicky, que cuando más he necesitado un respiro, siempre se ha ido conmigo a cualquier rincón del mundo a ver 'pedrolos' de los que me gustan, o pueblos con encanto, y así recargar pilas.

Y por supuesto, a una de las principales personas que ha hecho posible que consiga terminar esta tesis, mi director, pero sobre todo mi amigo, David. Empezaste siendo mi tutor de proyecto fin de carrera, casi allí en el pleistoceno, te convertiste después en mi amigo, y quien nos iba a decir que después de tanto tiempo me acabarías dirigiendo la tesis. Gracias por haber creído siempre en mí, en que podía lograrlo, incluso cuando yo he dudado. Y también dar las gracias a Soni, que durante todos estos años, también se convirtió en mi amiga, fuimos las mejores compis de trabajo, y siempre ha estado intentando ayudarme en todo.

Con mis compis de master Hector y Saúl, ahora amigos, empecé el camino de realizar esta tesis. Hemos vivido muchas etapas juntos, realizando cada uno nuestras tesis, y siempre ayudándonos, ¡muchas gracias chicos!. Pero en especial quería agradecer a Héctor, que siempre ha estado ahí, siendo el mejor compi en todos los trabajos hemos hecho juntos, escuchándome cuando lo necesitaba, y viviendo conmigo esa experiencia que fue la estancia en Kent. Gracias a ti aquello se convirtió no solo en trabajo, si no en visitar sitios alucinantes, descubriendo rosas rojas e historia por todos los rincones.

A todos mis compañeros del grupo AIDA, los que estáis día a día conmigo: Antonio, Alejandro, Cristian, Irene, Raquel, Raúl, Slavik, Victor y Patricia (que la hemos adoptado). Esos cafés, agobios compartidos, risas, y escapadas con vosotros hacen que el trabajo sea mucho más que eso. Pero en especial quiero dar las gracias a mis niños, Alejandro, Cristian y Victor por ser unos soles de personas, haberme ayudado tanto siempre, animándome, y haciendo que al final todo pareciera más fácil. Y aunque ya no esté en el grupo, también quería dar las gracias a Rus, por ser una de las mejores estudiantes que ha pasado por el grupo, pero sobre todo por haberme dejado conocer a una gran persona.

Además, quería agradecer a Julio Hernández Castro que me acogiera en la Universidad de Kent, dándome durante mi estancia allí la idea de utilizar las comunidades que hablaban sobre vacunación como dominio de aplicación.

Y por último, me gustaría dar las gracias al Departamento de Informática de la Universidad Autónoma de Madrid por haberme dado la oportunidad, y los recursos necesarios para realizar esta tesis.

Contents

List of Figures

List of Tables

List of Acronyms and Symbols

CDP	Community Detection Problem
CDA	Community Detection Algorithm
GA	Genetic Algorithm
SN	Social Networks
MOGA	Multi-Objetive Genetic Algorithm
CC	Clustering Coefficient
TPR	Triangle Participation Ratio
LCC	Local Clustering Coefficient
LWCC	Local Weighted Clustering Coefficient
GCC	Global Clustering Coefficient
D	Density
CN	Clique Number
C_D	Centralization
H	Heterogeneity
Exp	Expansion
Sep	Separability
CR	Cut Ratio
EBC	Edge Betweenness Centrality
CPM	Clique Percolation Method
CONGA	Cluster-Overlapping Newman Girvan Algorithm
CONGO	CONGA Optimized Algorithm
BIGCLAM	Cluster Affilitation Model for Big Networks
NMF	Non-negative Matrix Factorization
MDL	Minimum Description Length
CoDA	Communities through Directed Affiliations
EC	Evolutionary Computation
GP	Genetic Programming
EP	Evolutionary Programming
ES	Evolutionary Strategies
AI	Artificial Intelligence
NMI	Normalized Mutual Information
WHO	World Health Organization
ECDC	European Centre for Disease Prevention and Control
NLP	Natural Language Processes

GPHIN	Global Public Health Intelligence Network
SARS	Severe Acute Respiratory Syndrome
NER	Named-entity Recognition
VASSA	Vaccine Attitude Surveillance using Semantic Analysis
VAERS	Vaccine Adverse Event Reporting System
OCDP	Overlapping Community Detection Problem
GCF	Genetic-based Community Finding
MDF	Minimal Distance Fitness
MCCF	Maximum Clustering Coefficient Fitness
MWCCF	Maximum Weighted Clustering Coefficient Fitness
HF	Hybrid Fitness
CF	Centroid Fitness
MOGA-OCD	Multi-Objective Genetic Algorithms for Overlapping Communities Detection

Chapter 1

Introduction

> *"There are too many questions, there is not one solution, there is no resurrection, there is so much confusion."*
>
> - Madonna (Love Profusion - American Life)

This chapter presents the motivation and overview of this dissertation. Firstly, the Community Detection Problem (CDP) is introduced as main research focus of this work. Section 1.2 motivates the research questions that are addressed later. Section 1.3 briefly describes the Community Detection Algorithms (CDAs) and Genetic Algorithms (GAs), in order to provide a basic framework for the research questions that are described in Section 1.4. Next, the dissertation structure is described in Section 1.5 and, finally, the main contributions and the associated publications related to the this thesis are presented in Section 1.6.

1.1 The Community Detection Problem

The CDP can be defined as the division of a graph into clusters of nodes based on its topology information, where each cluster includes strongly interconnected nodes, and sparsely connections to the rest of the graph. This problem is similar to the idea of graph partitioning into groups of nodes according to the network topology. A partition is a division of a graph where each node belongs to only one community. Nowadays, this is an important problem applied in disciplines such as sociology, biology, neuroscience, or computer science, whose information can be easily represented using connected networks or graphs [58]. In addition, the emergence and exponential growing of the Social Networks (SNs) has led to a great interest in this topic, being currently one of the most popular and hot topics in computing research. Today, society lives in a connected world in which communication networks are intertwined with daily life, and social networks are one of the most important sources of social big data [21]. In social networks, individuals interact with one another and provide information on their preferences and relationships which can be easily represented as a graph. Therefore, these networks have become important sources for collective intelligence extraction, where community detection algorithms can be applied to discover useful knowledge on several fields.

The goal of the CDP is similar to the idea of graph partitioning in graph theory [40, 154]. A cluster in a graph can be mapped into a community. Therefore, this problem can be handled using graph-based clustering algorithms. The clustering problem can be described as a blind search on a collection of unlabelled data, where elements with similar features are grouped together in sets. There are three main techniques to deal with the clustering problem [93]: overlapping [25] (or non-exclusive), partitional [120] and hierarchical [108]. Overlapping clustering allows each element to belong to multiple clusters, partitional clustering consists in a disjoint division of the data where each element belongs only to a single cluster, and hierarchical clustering nests the clusters formed through a partitional clustering method creating bigger partitions, grouping the clusters by hierarchical levels.

In some domains, it could be interesting to allow that a node could belong to several clusters. For instance, it is well-known that people belonging to a social network can be member of multiple communities. A common feature observed in real-world networks is the existence of overlapping communities where a particular node can belong to more than one community. To solve this problem, fuzzy clustering algorithms applied to graphs [92] and overlapping approaches [178] have been proposed. In this thesis, the approach has been focused on the overlapping clustering techniques.

There is no single definition accepted to clearly state what is a cluster in a graph, and the variants used in the literature are numerous. But these kinds of algorithms are typically based on the topology information of the graph or network. Related to the graph connectivity, each cluster should be connected; it means that should exist paths connecting each pair of vertices within the cluster. It is generally accepted that a subset of vertices forms a good cluster if the induced sub-graph is dense, and there are few connections from the included vertices to the rest of the graph [98]. Considering both features, connectivity and density, a possible definition of a graph cluster could be a connected component where every two vertices in the sub-graph are connected by an edge [31].

This family of algorithms can be improved using Genetic Algorithms (GAs) to decrease their high computational complexity when they are applied to networks of very large sizes. A genetic algorithm is inspired by biological evolution [110], where the possible problem solutions are represented as individuals belonging to a population. The individuals are encoded using a set of chromosomes (called the genotype of the genome). Later these individuals are evolved, during a number of generations, following a survival/selection model where a fitness function is used to select the best individuals from each generation. Once the fittest individuals have been selected, the algorithm reproduces, crosses and mutates them trying to obtain new individuals (chromosomes) with better features than their parents. The new offspring and, depending on the algorithm definition, their parents, will pass to the following generation. This kind of algorithms have been usually employed in optimization problems [6, 75], where the fitness function tries to find the best solution among a population of possible solutions that are evolving. In other approaches, such as clustering, the encoding and optimization algorithm are used to look for the best set of groups that optimizes a particular feature of the data.

Several of these evolutionary clustering algorithms use a single optimization criteria as an objective function, such as modularity [161]. There are also GAs where the community detection is solved as a multiobjective optimization problem, generally optimizing two criteria [83, 143]. This thesis aims to carry out a comparative assessment of measures related to the topology of the network (Density, Triangle Partition Ratio, Clique Number, Clustering Coefficient, Separability, etc...), analysing their possible use as optimization criteria in order to find new approaches to improve CDAs based on GAs.

1.2 Motivation of the dissertation

The CDP has been the subject of many studies in the field of Data Mining [23, 137] and Social Network Analysis [20, 159]. Several methodologies have been applied to find optimal groups of nodes into communities. Usually these methods require a vast amount of memory and computational time to process large-scale networks in real world domains such as the World Wide Web, citation networks, social networks, or biochemical networks amongst others [58].

A major problem in graph clustering and CDPs is to look for a quantitative definition of a good community (or cluster), due to the definition often depends on the specific application domain. But, according to the nature of the considered problem, there must be more edges within each community than edges linking with the rest of the graph. There are several definitions related to what it means a *good community* within a graph in the literature. For this reason, a large number of different methodologies have been applied to solve this problem [154], where most of them are typically based on the topology information of the graph.

Other problem is derived from the fact that CDAs are based on the topology information of the graph to partition it. Due to this, usually these algorithms obtain goods results depending on the graph structure. For example, the algorithms based on the greedy optimization of modularity tends to form large communities rather than small ones, which often obtains poor values of modularity. Therefore this type of algorithms is not usually able to achieve good results for sparse graphs with small communities.

This dissertation aims to look for new evolutionary approaches that can aid to manage theses problems, providing new methods to find overlapping and stable communities within a graph. For this purpose several algorithms have been designed using different encodings, and fitness functions, to guide the searching process.

In addition, to evaluate the usefulness of the CDA algorithms in a real domain, this dissertation also includes the application of them to identify opinion communities related to Public Healthcare topics in Social Networks. 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 [95, 136, 169].

In the related literature, there are several works investigating knowledge acquisition from social networks about vaccine sentiments using classification techniques [33, 113, 153]. 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 do 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. For this reason, this dissertation proposes the use of CDAs, which do not require that human-labelling process, to identity communities in Twitter which are disseminating vaccine opinions. Then, these communities can be analysed using different network metrics to identify the most relevant users, and to analyse how them can influence to the rest of users in a particular community, zone, or country.

1.3 Problem statement

The clustering problem can be defined as a blind search on a dataset whose objective is to partition the dataset into groups of elements sharing similar features. Some classical solutions such as K-means (for a fixed number of clusters) [120] or Expectation-Maximization [56] (for a variable number of clusters), amongst others, are based on distances or metrics that are used to determine the similarity between the elements of the dataset.

When the dataset is represented as a graph (such as in the CDPs), the clustering problem is called graph clustering, and it consists of the grouping of the nodes belonging to this graph into clusters. Therefore, the aim of the CDAs is similar to the graph clustering. Graph models are useful for diverse types of data representation, and nowadays they have become especially popular, being widely applied in the social networks area. Graph models can be naturally used in these domains, where each node or vertex can be used to represent an agent, and each edge is used to represent their interactions. Later, algorithms, methods and graph theory have been used to analyse different aspects of the network, such as: structure, behaviour, stability or even community evolution inside the graph [55, 73, 128, 175]. Depending on the methodology applied to perform the graph partitioning, there are several different methods for graph clustering such as cutting, multi-level partitioning, spectral analysis, or connectivity analysis, amongst others [156].

GAs [90] have been traditionally used in a large number of different domains, mainly related to optimization problems [26, 39, 51, 65, 155]. This kind of algorithms is one of the most successful algorithm in the Evolutionary Computation (EC) area. The EC is a field of AI, inspired in the biological evolution where all the EC techniques share three characteristics: (1) they use a population of individuals that represent solutions, (2) individuals are modified using at least one genetic operator, and (3), individuals are under a selective pressure, where the fittest ones can reproduce their characteristics. The result is an evolving population of partial solutions that would eventually converge to a global solution.

GAs have demonstrated to be robust algorithms able to find satisfactory solutions in problems with large search space evaluating a minimum number of points. In particular to solve *graph clustering* problems, GAs have been used to look for the best partition of the graph in groups of nodes (clusters) optimizing a particular feature of the graph structure.

Regarding to the measures used to guide the algorithms finding optimal clusters, there are two main approaches [61]: the first one computes some values from the vertices and then classify them into clusters; the second one compute a fitness measure over the set of possible clusters choosing the optimal among all cluster candidates. For the first approach, there are several similarity, or distance, metrics that can be applied such as the Euclidean Distance, the Jaccard index or Cosine similarity, among others [119]. The other approaches use connectivity measures computed over the set of possible graph clusters such as density, clustering coefficient or clique number, as fitness function of the algorithm [173].

Taking into account the different existing methodologies, the first algorithms designed and implemented in this thesis are based on the idea of combining classical clustering methods with the graph clustering methods based on connectivity measures (K-fixed and K-adaptive GCF-I). These new algorithms adopt an evolutionary approach to optimize the search of the communities, using different fitness functions that combine distances metrics based on node features, with measures related to the network topology of the graph.

Both GCF-I algorithms require that each node has a set of features associated, which allow to measure its distance or similarity related to the rest of nodes belonging to the graph. However, many dataset collections representing graphs do not have features associated to their nodes. Therefore, a new version of the algorithm (GCF-II) was implemented where the graph partitioning is only performed optimizing measures based on the own network topology of the graph. For this purpose, a new fitness function has been implemented combining several measures extracted from Graph Theory (Density, Centralization, Heterogeneity, Neighbourhood, Clustering Coefficient).

Once the first generation of algorithms are experimentally tested, it can be noticed that they present the same problem as the CDAs of the State of the Art. These algorithms show a dependence on the graph structure. In order to handle with this problem, these algorithms are extended to multi-objective approaches trying to promote solution diversity, in a second generation of previous algorithms (MOGA-OCD). The fitness function designed for this proposed approach optimizes simultaneously two objective functions: one is used to maximize the internal connectivity of the node groups, whereas the second one tries to minimize the external connections to the rest of the nodes. For this purpose, several internal and external connectivity measures from the graph theory, are used as the optimization criteria of the algorithms. Since many connectivity measures can be used as the objective functions, a comparative assessment of these metrics has been carried out in this thesis. This study allow to identify the most suitable measures for partitioning the graph into overlapping communities.

Finally, the last part of this thesis has been focused on the application of the CDAs to identify opinion communities related to Public Healthcare topics in Social Networks. Both, classical algorithms and the new algorithms implemented, have been used to detect communities in Twitter that are talking about vaccines. Finally, an experimental analysis of the influence of these communities to the rest of users, in a particular zone or country, is carried out proving how useful is this new knowledge to Public Healthcare Organizations. With these information, these organizations could improve their strategies, increasing control and preventive measures in the risk zones identified.

1.4 Research Questions

This PhD Thesis aims to analyse, and study, the possible combination of graph clustering methods and GAs in order to understand in depth the process of these algorithms, and to find new approaches to improve the classical techniques for solving the CDP. In addition, a detailed study of the most appropriate measures related to the topology structure of the networks has been carried out. This fact allows to select the most suitable metrics such as objective functions to guide the search of the GAs. To achieve these main objectives the main research questions of this thesis can be described as follows:

- **Q1:** Is it possible to combine clustering methods based on distances with genetic graph-based approaches to improve the results obtained by classical overlapping community detection algorithms?

- **Q2:** Are these algorithms able to identify quality communities only using network topology measures as objective functions?
- **Q3:** How can multi-objective genetic algorithms deal with the problem of dependence on the graph structure shown by the classic algorithms?
- **Q4:** How can these algorithms be applied to real social domains to acquire useful information?

1.5 Structure of the thesis

The dissertation has been structured in six chapters. A brief description of the chapter contents are given as follows:

- **Chapter 1: Introduction.** It provides a general context and motivations related to the dissertation. In addition, the main objectives and research questions are introduced, as well as the main contributions and publications generated in this thesis.
- **Chapter 2: State of the Art.** It introduces the State of the Art related to Community Detection Algorithms and Genetic Algorithms in order to contextualize the different contributions of this PhD thesis. For this purpose, firstly some basic definitions from Graph Theory are presented. Finally, some current applications using information extracted from SNs are described as possible real domains of application.
- **Chapter 3: Genetic Graph-based Approaches for Overlapping Community Detection.** This chapter presents the first algorithms developed based on a Genetic Algorithms and Graph Theory to solve Overlapping Community Detection Problems (OCDPs). These new algorithms optimize the search of the communities using different fitness functions, which combine distance metrics based on features of the nodes, with measures related to the network topology of the graph. Here, the analysis is focused on the different encodings and fitness functions which can deal with the CDP. Finally, all the new algorithms implemented have been experimentally evaluated using the Eurovision Contest dataset, that can be considered as a well-known SN.
- **Chapter 4: Multi-Objective Genetic Approaches for Overlapping Community Detection.** It describes the second generation of algorithms based on a multi-objective approach, that extends the previous generation to deal with the problem of the dependence on the graph structure. Two different encodings are designed, one based on nodes whereas the other one is based on edges, to evaluate the two different approaches that are current in use for finding communities. Finally, the proposed algorithms have been compared among themselves, and evaluated against other well-known algorithms from the state of the art.
- **Chapter 5: Applications for detecting communities on Social Graph-based Information.** This chapter presents the application of the Community Detection Algorithms designed in a real domain using social graph-based information. In this particular case, the algorithms are applied to detect communities in Twitter that are disseminating opinions about vaccination. In addition, an analysis of the influence of these communities to the rest of users, in a particular zone or country, is carried out. This provides an idea of how

useful can be this new knowledge for the Public Healthcare Organizations, to improve their immunization strategies, or even to apply preventive measures in the risk zones identified.

- **Chapter 6: Conclusions and Future Work.** The Research Questions described in Chapter 1 are addressed in order to provide some answers, based on the results obtained from this research. Finally, taking into account all the analysis carried out, a summarize of the possible future works is presented.

1.6 Publications and Contributions

This Section shows the contributions that have been generated during the development of this thesis. These publications have been organized by journals and conferences, and sorted by year.

International Journals

(IJ-1) Gema Bello-Orgaz, Hector D. Menendez, David Camacho: *Adaptive K-Means Algorithm for overlapped graph clustering.* International Journal of Neural Systems. Ed. By World Scientific. Vol. 22 Issue 5, pp. 1-19, 2012.
Impact factor = 5.054 (JCR, 2012) [Q1].

- Contribution: This contribution is related to Section 3.1.

(IJ-2) Gema Bello-Orgaz, Hector D. Menendez, Shintaro Okazaki, David Camacho: *Combining Social-based Data Mining Techniques to Extract Collective Knowledge from Twitter.* Malaysian Journal of Computer Science. Ed. By World Scientific. Vol. 27, Issue 2, pp. 95-111, 2014.
Impact factor = 0.405 (JCR, 2014) [Q4].

- Contribution: This contribution is related to Section 2.3.

(IJ-3) Gema Bello-Orgaz, Jason J. Jung, David Camacho: *Social big data: Recent achievements and new challenges.* Information Fusion, Ed. By Elsevier. Volume 28, pp. 45–59, 2016.
Impact factor = 4.353 (JCR, 2015) [Q1].

- Contribution: This contribution is related to Sections 2.2 and 2.3.

(IJ-4) Gema Bello-Orgaz, Julio Cesar Hernandez-Castro, David Camacho: *Detecting Discussion Communities on Vaccination in Twitter.* Future Generation Computer Systems, Ed. By Elsevier, Vol. 66, pp. 125-136, 2017.
Impact factor = 2.43 (JCR, 2015) [Q1].

- Contribution: The content of this paper is directly related to Chapter 5.

International Conferences

(IC-1) Gema Bello, Raul Cajias, David Camacho: *A Study on the Impact of Crowd-Based Voting Schemes in the 'Eurovision' European Contest.* In proceedings of 1st International Conference on Web Intelligence, Mining and Semantics (WIMS'11). ACM press. Sogndal, Norway, 25-27 May 2011.

- Contribution: This contribution is related to Sections 3.1.3 and 3.3.

(IC-2) Gema Bello, Hector D. Menendez and David Camacho: *Using the Clustering Coefficient to guide a Genetic-based Community Finding Algorithm.* In Proceedings of the 12th International Conference on Data Extraction and Automated Learning (IDEAL 2011). Lecture Notes in Computer Science (LNCS), Vol. 6936, pp. 160-169. Springer Berlin / Heidelberg. Norwick, UK, September 2011. **Core-ERA C**.

- Contribution: This contribution is related to Section 3.1.1.

(IC-3) Gema Bello, Hector Menendez, Shintaro Okazaki and David Camacho: *Extracting Collective Trends from Twitter using Social-based Data Mining.* In Proceedings of the 5th International Conference on Computational Collective Intelligence (ICCCI 2013). Ed. by Springer-Verlag, Vol. 8083, pp. 622-630. Craiova, Romania, September 11-13, 2013. **Core-ERA C**.

- Contribution: This contribution is related to Section 3.1.1.

(IC-4) Gema Bello-Orgaz and David Camacho: *Evolutionary clustering algorithm for community detection using graph-based information.* In proceedings of 2014 IEEE Congress on Evolutionary Computation (IEEE CEC). pp. 930 – 937. Beijing, China, 6-11 July 2014. **Core-ERA B**.

- Contribution: This contribution is related to Section 3.2.

(IC-5) Gema Bello-Orgaz, Julio Hernandez-Castro, and David Camacho: *A survey of social web mining applications for disease outbreak detection.* IDC 2014: 8th International Symposium on Intelligent Distributed Computing, Vol. 570, pp. 331-340. September 3-5, 2014 Madrid, Spain.

- Contribution: This contribution is related to Section 2.3.2.

Submitted International Journals

(SIJ-1) Gema Bello-Orgaz, Sancho Salcedo, David Camacho. *Multi-Objective Genetic Algorithm for overlapping community detection using graph-based information.* Swarm and Evolutionary Computation –SWEVO. Submitted 2017.
Impact factor = 2.963 (JCR, 2015) [Q1].

- Contribution: This contribution is related to Chapter 4.

Chapter 2

State of the Art

"To understand the things that are at our door is the best preparation"
for understanding those that lie beyond."

- Hypatia of Alexandria

This chapter starts with a general introduction to some basic definitions from Graph Theory which are usually used in graph clustering techniques. After this brief introduction, an overview of community detection algorithms is presented. Later, genetic algorithms are introduced showing how they can be applied to improve these techniques. Finally, some current applications to Social Networks of these type of data mining methods are described.

2.1 Basic Definitions of Graph Theory

Graph theory is an area of important contribution for research in data analysis. Graph models are useful to represent diverse types of data. They have become especially popular in the last years, and have been widely applied in the Social Networks (SN) area. These models can be naturally applied in these domains, where each node, or vertex, can be used to represent an user, and each edge is used to represent their interactions. Later, algorithms, methods and graph theory have been used to analyse different characteristics of the network, such as: structure, behaviour, stability or even community evolution inside the graph [55, 73, 128, 175]. Most of the community detection algorithms use concepts and metrics extracted from graph theory in order to split a graph into clusters or communities. For this reason, and before describing the different existing approaches in the state of the art, some of the basic concepts from graph theory are briefly introduced.

Definition 2.1.1 (Graph)**.** A graph $G = (V, E)$ is a set of vertices or nodes V denoted by $\{v_1, \ldots, v_n\}$ and a set of edges E where each edge is denoted by e_{ij} if there is a connection between the vertices v_i and v_j.

Graphs can be directed or undirected. If all edges satisfy the equality $\forall i, j,\ e_{ij} = e_{ji}$, the graph will be undirected.

The most usual approach to represent a graph is through its adjacency matrix which can be defined as:

Definition 2.1.2 (Adjacency Matrix)**.** An adjacency matrix of G, A_G, is a square $n \times n$ matrix where each coefficient satisfies:

$$(a_{ij}) = \begin{cases} 1, & \text{if } e_{ij} \in E \\ 0, & \text{otherwise} \end{cases}$$

When it is necessary to work with weights in the edges, a new kind of graph needs to be defined:

Definition 2.1.3 (Weighted Graph)**.** G is a weighted graph if exists a function $w : E \to R$ which assigns a real value to each edge.

Any algorithm that works with the vertices of a graph needs to analyse each node neighbours. The neighbourhood of a node is defined as follows:

Definition 2.1.4 (Neighbourhood)**.** Taking two edges, $e_{ij} \in E$ and $e_{ji} \in E$, we say that v_j is a neighbour of v_i. The neighbourhood of v_i Γ_{v_i} is defined as $\Gamma_{v_i} = \{v_j \mid e_{ij} \in E \text{ and } e_{ji} \in E\}$. Then, the number of neighbours of a vertex v_i is $k_i = |\Gamma_{v_i}|$

Once the most general and simple concepts from graph theory are defined, we can proceed with the definition of some basic metrics, based on previous definitions, which will be later used in the algorithms designed in this thesis. There are two basic types of metrics that can be used to provide a quantitative measure of the goodness for a graph partition [182]: ***internal*** community connectivity, and ***external*** connectivity of the community related to the rest of the network. A 'good' community should be cohesive, compact, and *well connected internally.* On the other hand, and regarding to external connectivity, it should be well *separated* from the rest of network nodes. According to this classification of connectivity types, the connectivity metrics used in this thesis are described below.

2.1.1 Metrics based on internal connectivity

1. **Triangle Participation Ratio (TPR)** [182]:

   This is a measure of graph cohesion, which is defined as the fraction of nodes in a graph that belongs to a triangle. A triangle, in graph theory, is a planar undirected graph with 3 vertices and 3 edges generating a complete graph in the form of a triangle, as shown in Figure 2.1.

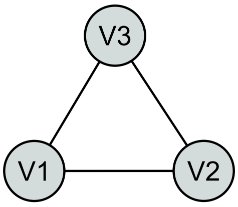


Figure 2.1: A triangle example into a graph.

Taking this into account, the TPR can be defined as follows:

$$TPR = \frac{|\{u : u \in V, \{(v,w) : v,w \in V, (u,v) \in E, (u,w) \in E, (v,w) \in E\} \neq \emptyset\}|}{|V|} \tag{2.1}$$

This ratio can take values from 0 to 1. If every node belongs to a triangle, the TPR value is equal to 1. Otherwise its value will be 0.

2. **Local Clustering Coefficient (LCC)**[55, 176]:

This coefficient measures the transitivity of a node into the graph, and it is usually used for undirected graphs to represent the probability that two neighbours of a vertex are connected. The transitivity property in a graph can be observed when triangles are formed. As was defined by Watts and Strogatz [176], suppose that a vertex v_i has k_i neighbours; then a maximum of $k_i(k_i-1)/2$ edges can exist between them. This happens when every neighbour of v_i is connected to every other neighbour of v_i. Taking this into account, the LCC of a node v_i is defined as the fraction of the number of connected pairs between all neighbours of v_i, and this maximum possible number of edges between all neighbours:

$$LCC_i = \frac{2 \times \sum_{j,h} a_{jh}a_{ij}a_{ih}a_{ji}a_{hi}}{k_i(k_i-1)} \tag{2.2}$$

The LCC measure provides values ranging from 0 to 1. Where 0 means that the node and its neighbours do not have clustering features, so they do not share connections between them. Whereas, the value 1 means that they are completely connected, as shown in Figure 2.2.

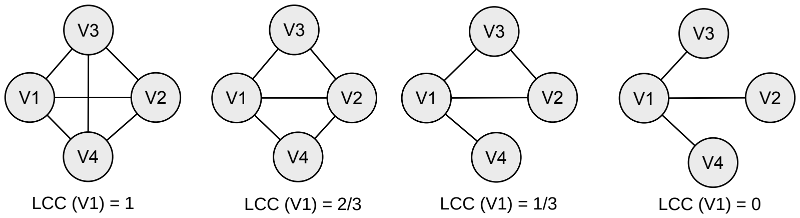


Figure 2.2: Examples of Local Clustering Coefficient for the node V1 showing different possible graphs, and their related LCC value.

3. **Local Weighted Clustering Coefficient (LWCC)**[15]:

To study weighted graphs, the definition of LCC can be extended. Following the same assumption of LCC definition, let W be the weight matrix with coefficients w_{ij} and A be the adjacency matrix with coefficients a_{ij}, if we define:

$$S_i = \sum_{j=1}^{|V|} a_{ij}a_{ji}w_{ij} \tag{2.3}$$

Then, the Local Weighted Clustering Coefficient can be defined as :

$$LCC_i^w = \frac{2 \times \sum_{j,h} \frac{(w_{ij}+w_{ih})}{2} a_{jh}a_{ij}a_{ih}a_{ji}a_{hi}}{S_i(k_i-1)} \tag{2.4}$$

For this new definition, the connections between the neighbours of a particular node are considering, but now adding the weight information related to the original node. This new measure calculates the distribution of the weights of the node that we are analysing, and shows how good the connections of that cluster are. The LWCC has the same value than the LCC when all the weights are fixed to the same value

4. **Global Clustering Coefficient (GCC)** [176]:

   This coefficient measures the global transitivity of the graph providing a general overview of the graph structure. GCC is defined as the ratio of the triangles and connected triplets in the graph:

   $$GCC = \frac{3 \times |Triangles|}{|Triples|} \tag{2.5}$$

   It provides values from 0 to 1. If all possible connections among the neighbours of all the nodes into the graph are available, GCC gives a value of 1. A network with GCC close to 1 contains highly connected clusters. Otherwise, if there are no connections between the neighbours nodes, this coefficient has a value of 0.

5. **Density(D)** [174]:

   This is defined by the number of edges in the graph G divided by the total number of possible edges, it can be expressed as follows:

   $$D = \frac{m}{n(n-1)/2} \tag{2.6}$$

   where the number of the nodes within the graph is $n = |V|$, and the number of edges is $m = |E|$.

   The density is a real value between 0 and 1. Any graph that does not contain any edge, so all the nodes will be isolated, will have a density of 0, whereas for a full connected graph, where every node is connected to the rest of the nodes in the network, will have a density of 1.

6. **Clique Number (CN)** [10]: A *clique* of a graph is a subset of mutually adjacent vertices in V (every two vertices in the subset are connected by an edge). It means that the induced subgraph from these nodes is complete. A clique is called **maximal** if it is not contained in any other clique, as shown in Figure 2.3. The size of the maximum clique is called the clique number of the graph that is denoted by $\omega(G)$ [10].

   The Maximum Clique Problem is one of the classic NP-complete problems, and there are several proposed algorithms in the literature to manage it. The Bron-Kerbosch algorithm [34] is one of the most well-known and widely used method based on a recursive backtracking search. Tomita et al. [167] proposed a similar technique to the Bron-Kerbosch algorithm using a depth-first search algorithm with pruning methods. Finally, in the last years, different variations of the Bron-Kerbosch algorithm have been implemented to be applied into larger graphs as the method presented by Eppstein et al. [62].

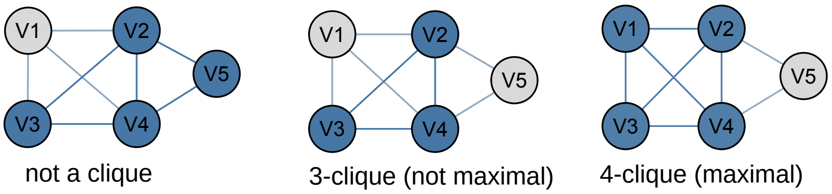


Figure 2.3: Examples of different cliques contained in a graph containing 5 nodes.

7. **Centralization** (C_D) [74]:

This metric is based on the concept of degree centrality of the nodes of a graph, which is defined as the number of edges incident upon of them. According to this, the degree centrality ($C_d(v_i)$) of a vertex v_i for a given indirected graph G is defined as its neighbourhood (k_i) (see Figure 2.4). In the case of a directed graphs, it is possible to define two distinct measures of degree centrality; namely **indegree** and **outdegree**. Accordingly, indegree centrality corresponds to the number of links directed to the node, otherwise outdegree centrality is the number of links that the node directs to the rest of nodes of the graph.

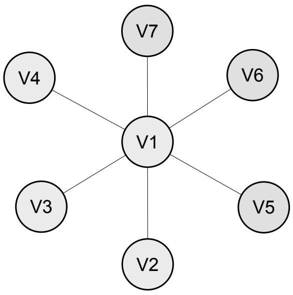


Figure 2.4: Example of degree centrality. In this graph, the degree centrality for the $V1$ node is $C_d(V1) = 6$, whereas for the rest of nodes, this value is 1.

The definition of degree centrality on the node level can be extended to the whole graph. Freeman [74] provided a measure of graph centralization based on differences for the node centralities. This index of graph centralization has two main features: (1) it should index the degree to which the centrality of the most central node exceeds the centrality of all other nodes, and (2) it should be expressed as a ratio of that excess to the maximum possible value for a graph containing the same number of nodes. The maximum difference for node centralities in a graph takes place when the graph contains one central node to which all other nodes are connected (a star graph), as shown in Figure 2.4. In this case the difference has a value equal to $n^2 - 3n + 2$ where $n = |V|$. Thus, let $v*$ be the node with highest degree centrality, the degree centralization of the graph G is defined as follows:

$$C_d(G) = \frac{\sum_{i=1}^{n} [C_d(v*) - C_d(v_i)]}{n^2 - 3n + 2} \tag{2.7}$$

As previously mentioned, Freeman proved that this measure takes its maximum value, 1, for those graphs whose topology is a star or a wheel, whereas decentralized graphs are characterized by having a centralization close to 0.

8. **Heterogeneity (H)** [57]:

   The heterogeneity of the degree distribution has been the focus of considerable research in recent years. Many measures of network heterogeneity are based on the variance of the connectivity. This measure notices the tendency of a network to contain "hub" nodes. Dong and Horvath [57] defined this measure as the coefficient of variation of the connectivity distribution:

$$H = \frac{\sqrt{variance(k)}}{mean(k)} \tag{2.8}$$

2.1.2 Metrics based on external connectivity

1. **Expansion (Exp)** [182]:

   This metric measures the average number of external edges per node in a community of the graph. Figure 2.5 shows an example of a community containing 4 nodes (market in dark blue) where it can be appreciated the external edges for this community, which are connecting the nodes belonging to this community, to others outside of it (marked in bold).

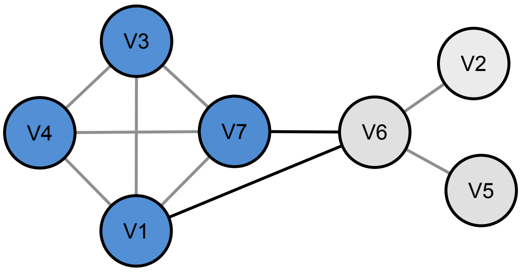


Figure 2.5: Example of external edges of a community belonging to a graph. The nodes V1, V3, V4 and V7 belong to a community where the external edges of it are marked in bold.

   Then, let C be a subset of nodes in the graph, the expansion of C is defined as the ratio between the number of external (boundary) edges in the subset, and the number of nodes belonging to it:

$$Exp(C) = \frac{|\{(u,v) \in E : u \in C, v \notin C\}|}{|C|} \tag{2.9}$$

2. **Separability (Sep)** [182]:

   Separability measure is used to represent how communities are separated from the rest of the graph. This metric is the ratio between the number of internal edges and external edges in a community. Figure 2.6 presents an example of a graph containing a community where the internal edges belonging to it are marked in bold. Following the same example, in previous Figure 2.5, it can be viewed which are the external edges of this community.

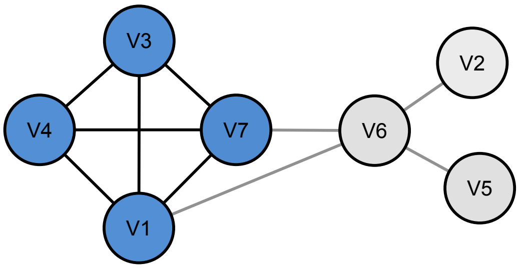


Figure 2.6: Example of internal edges of a community belonging to a graph. The nodes V1, V3, V4 and V7 belong to a community where the internal edges of it are marked in bold.

So, let C be a subset of nodes in the graph, the Sep value of this subset is defined as follows:

$$Sep(C) = \frac{|\{(u,v) \in E : u \in C, v \in C\}|}{|\{(u,v) \in E : u \in C, v \notin C\}|} \tag{2.10}$$

3. **Cut Ratio (CR)** [182]

   This metric provides the ratio between the number of external edges in a community (see Figure 2.5) and all the possible edges leaving the community. If n_c represents the number of the nodes in the community ($n_c = |C|$), and n is the number of nodes in the graph ($n = |V|$), the maximal possible number of external edges of a particular community will be $n_c(n - n_c)$. Taking this into account the CR is defined as follows:

$$CR(C) = \frac{|\{(u,v) \in E : u \in C, v \notin C\}|}{n_c(n - n_c)} \tag{2.11}$$

2.2 Community Detection Algorithms

The main goal of graph partitioning on Graph Theory [40, 154] is similar to the idea of Community Detection Problem (CDP). In computer science, the process of identifying the underlying structure of the data in terms of grouping the elements is called clustering [94], and a cluster in a graph can be called community. Therefore, graph clustering [61, 124] is the task of grouping the vertices into clusters or communities considering the edge structure of the graph. A complete roadmap of graph clustering can be found in the paper carried out by Schaeffer [156], where different clustering methods are described and compared using different types of graphs: weighted, directed and undirected. These methods are: cutting, spectral analysis and degree connectivity (an exhaustive analysis of connectivity methods can be found in Hartuv and Shamir work [89]) amongst others. This roadmap also provides an overview of computational complexity from a theoretical and experimental point of view for the studied methods.

There are several definitions related to what it means a *good community* within a graph in the literature. For this reason a large number of different methodologies have been applied to solve this problem [154], and most of them are typically based on the topology information of the graph. However, usually there are two main approaches to deal with the CDP [93]: ***partitional*** and ***overlapping*** (or non-exclusive).

This section starts with a description of the main measures usually applied to evaluate the quality of the resulting communities detected by the CDAs. Then, a general introduction of the partitional techniques, which perform a disjoint division of the data where each element belongs only to a single community, are introduced. Following, an overview of overlapping methods that allow to each node belonging to multiple communities is described. And finally, some relevant ***evolutionary approaches*** that try to improve the computational complexity of previous methods are described.

2.2.1 Evaluation Metrics

There are two main measures frequently used to evaluate the effectiveness of the community detection algorithms; the **Modularity** and the **Normalized Mutual Information (NMI)**. Modularity is one of the most popular quality function to assess a graph partition. It is based on the computing of the density for the subgraphs contained into this partition, and it can be used for any dataset that can be represented as a graph. On the other hand, NMI is a metric related to the Information Theory that can be used to calculate the similarity between two graph partitions. Therefore, applying NMI to a dataset with ground-truth (the dataset has been pre-tagged by a human expert), it is possible to evaluate the accuracy of the outcomes achieved by a specific algorithm. However, the standard definitions of these measures are only used to compare disjoint partitions, and the algorithms mentioned in this thesis have been designed to detect overlapping communities. So it is necessary to used extensions of these standard measures for partitions with overlaps:

- **Modularity (Q)**: The modularity given by Newman and Girvan [80] is a well-known quality measure to evaluate the goodness of a partition. This metric is based on the idea that a subgraph can be considered modular, or a community, if the density of the subgraph is higher than that expected for a corresponding random subgraph. Therefore, to calculate the modularity for a graph partition, the subgraphs contained in it are compared against a random subgraphs with the same number of nodes, the same number of edges, and the same degree distribution as in the original subgraph, but with the edges among nodes randomly generated. This is called the called 'null-model'.

  Nicosia et al. [134] extend this definition taking into account that each node can belongs to many communities with a certain strength for each one. Therefore, in this new definition of the modularity will be considered a set of "belonging factors" where each factor represents how strongly a node i belongs to a community c. In addition, a coefficient belonging to each community for edges incoming to, or outgoing from, a node is defined. The definition of this function is somewhat arbitrary, for example, it could be define as the product of the belonging coefficients of the nodes involved, or the maximum coefficient, etc... Using both new concepts, the modularity is reformulated according to a given community can be weighted by the corresponding belonging coefficients. As in the standard definition, a higher positive value of this extended Modularity corresponds to a more modular partition, being a better partition of the graph in terms of quality. A result close to 0 is given when the communities detected are similar to a randomly partition of the graph. And finally, higher negative values means worst partitions without structure.

- **Normalized Mutual Information**: When the ground truth of the community structure is known for a dataset, the Normalized Mutual Information (NMI) [50] can be applied to measure the similarity between partitions. This allows to compare covers, which in this specific case is the division of a network into communities detected by a specific algorithm, against the division given in the ground-truth of the dataset.

The NMI definition was extended by Lancichinetti et al. [109] to the the case of overlapping communities. Mutual information is a measure related to the inherent dependence expressed in the joint distribution of X and Y relative to the joint distribution of X and Y under the assumption of independence. It can be expressed using the marginal entropies of X ($H(X)$) and Y ($H(Y)$). In this extension for overlapping partition, the membership of the node is considered as a binary array with a length equal to the number of communities of the partition. In this array, each element represents if the node belongs to a particular community or not. Taking into account this information, the joint distribution are reformulated to consider all possible belongings of a node to different communities. The value of NMI will be 1 when the clusters are fully matched with the ground truth.

2.2.2 Partitional Methods

In partitional methods a disjoint division of the graph is performed, so each vertex (a graph node) will only belong to a single community. One of the most well-known partitional algorithms for community detection is the Edge Betweenness Centrality (EBC) proposed by Girvan and Newman [133]. This method uses a similarity measure called "***edge betweenness***" based on the number of the shortest paths between all vertex pairs. The proposed algorithm iteratively removes links with the highest betweenness score (edges lying between communities) to achieve the isolation of the communities. The *EBC* algorithm can be summarized as shown in Algorithm 1.

Algorithm 1: Edge Betweenness Centrality Algorithm

```
Input: A graph $G = (V, E)$ where $V$ is a set of vertices denoted by $\{v_1, \ldots, v_n\}$ and $E$ is
       a set of edges $E$ denoted by $e_{ij}$ representing whether there is a connection
       between the vertices $v_i$ and $v_j$.
Output: Communities extracted from the Dendrogram of removals (D)
while $E \neq \emptyset$ do
    $B \leftarrow calculateBetweenness(E)$
    $e_{ij} \leftarrow selectEdgeHighest(B, E)$
    $E \leftarrow E - e_{ij}$
    $D \leftarrow dendrogramRemovals(e_{ij}))$
return $D$
```

Figure 2.7 shows how EBC algorithm will partition a graph following the process previously described. In this example, the graph is partitioned into three communities removing three edges. In the first step, the edges which have higher betweenness value are $e(4,5)$ and $e(4,6)$, thus one of them should be removed. As shown this figure the first edge removed is $e(4,6)$. In the second iteration of the algorithm, the edge with higher betweenness will be $e(4,5)$, which will be removed. Finally, computing all the edge betweenness in the third iteration, the edge $e(7,9)$ will obtain the highest betweenness value, and the process is finished. The final result after applying the EBC algorithm is a top-down dendrogram where the input graph is split up into different communities. As shown in Figure 2.7 the three resulting communities of this example are: $\{1, 2, 3, 4\}\{5, 6, 7, 8\}\{9\}$.

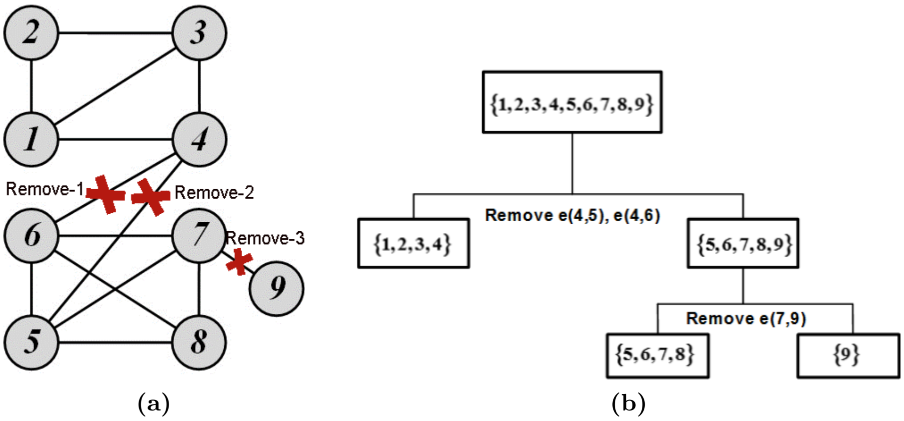


Figure 2.7: Example of EBC algorithm execution. It has been taken from [184].

On the other hand, many authors have been employed the *Modularity* as basis of their algorithms to identify community structures into networks. As described in Section 2.2.1, this is the most popular evaluation metric to measure the quality of a graph partition, and it is frequently used on graph clustering techniques. The algorithms based on this metric usually show and excellent performance when the size of the network is small [41, 131]. Updating the modularity matrix in the Newman algorithm involves a large number of useless operations, due to the sparsity of the adjacency matrix. This was improved by Clauset et al. [40] work using the matrix of modularity variations to perform more efficiently. But, the major disadvantage of these types of algorithms is still the high computational complexity of this metric on networks of very large sizes.

For this reason, this modularity measure has been modified trying to reduce the computational demands significantly through several new approaches. Currently, there are several algorithms based on good approximations of the modularity which are able to detect communities in a reasonable time. The first technique based on a greedy optimization of modularity was proposed by Newman [132] (called ***fastgreedy***). It is an agglomerative hierarchical clustering algorithm, where groups of vertices are successively joined to form larger communities such that modularity increases after the merging.

The greedy optimization of modularity tends to form large communities rather than small ones which often obtains poor values of modularity. In order to solve this problem, Danon et al. [49] proposed to normalize the modularity variation in order to improve the identification of small communities. This approach obtains better results especially when communities are very different in size. Another approach to avoid the formation of large communities was proposed by Schuetz and Caflisch [157] merging more community pairs at each iteration. Moreover this modification of the greedy algorithm is combined with a procedure, where single vertices are moved to the neighboring community, that yields the maximum increase of modularity.

Despite the improvements and modifications of the accuracy of these greedy algorithms, they usually do not obtain good results. For this reason, Newman reformulated the modularity measure in terms of eigenvectors by replacing the Laplacian matrix with the modularity matrix named spectral optimization of modularity [131]. This improvement have been also applied to improve the results of other optimization techniques [150, 172].

Finally other well-known algorithm, based on modularity, is the Multi-Level [28]. 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.

Random walks [152, 146, 187] can also be useful for finding communities. If a graph has a strong community structure, a random walker spends a long time inside a community because of the high density of internal edges and the consequent number of paths that could be followed. Therefore, there are several algorithms to detect communities in the literature using this approach. Walktrap algorithm [146] is based on the idea that short random walks tend to stay in the same community. This work proposed a new measure of similarity between vertices based on random walks which can be computed efficiently, and and it is used in an efficient hierarchical agglomerative algorithm that detects communities in a network. Other approach was propoed by Zhou and Lipowsky [187], based on the fact that walkers move preferentially towards vertices that share a large number of neighbours. They defined a proximity index that indicates how close a pair of vertices is to the rest of vertices in the graph. Then communities are detected using a procedure called NetWalk, which is an agglomerative hierarchical clustering method where the similarity between vertices is expressed by their proximity.

Finally, another approach based on random walks is the Infomap algorithm proposed by Rosvall et al. [152]. It 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 attempting to find a partition that yields the minimum description length of an infinite random walk on the network. The result is a map that simplifies, and highlights, the regularities in the structure, and their relationships as shown in Figure 2.8.

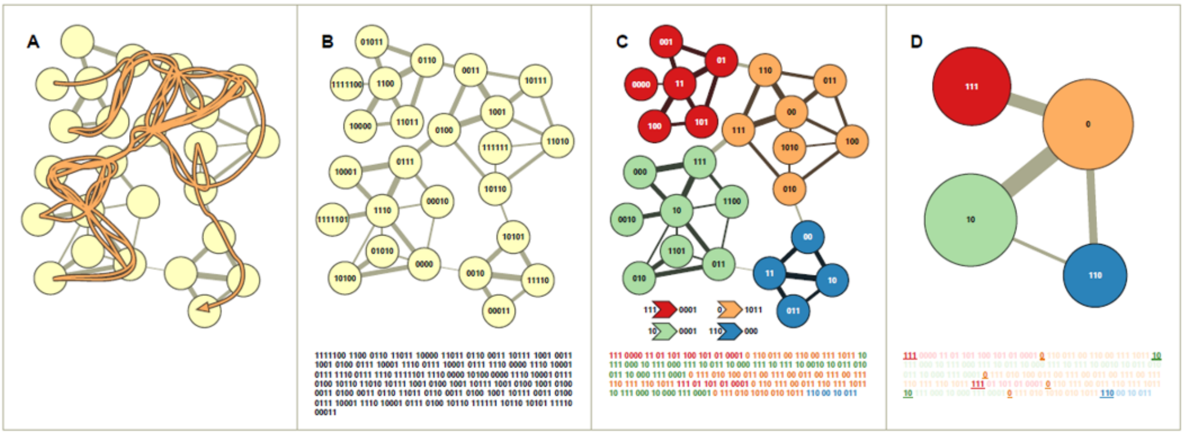


Figure 2.8: Example of the Infomap algorithm detecting communities by compressing the description of information flows on networks. This example has been taken from [152].

Another approach based on the ***Label Propagation*** of the vertex was described by Raghavan et al. [147]. In this new algorithm, initially each vertex is assigned to a different label. Then, all the vertices are iteratively labeled with the dominant label of its neighbourhood (see Figure 2.9). This process generates densely connected groups, which form a consensus with an unique label. The algorithm ends when all vertices reach a consensus.

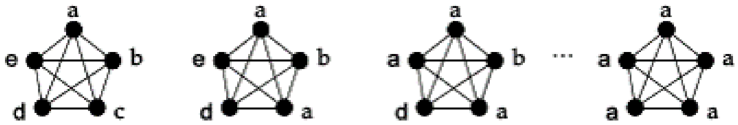


Figure 2.9: Example of Label Propagation algorithm [147], where nodes are updated one by one from left to right. In this case, and due to a high density of edges, finally all nodes acquire the same label.

Based on Potts ***statistical model***, Reichardt and Bornholdt proposed a method to detect communities. This method maps the graph onto a zero-temperature q-Potts model with nearest-neighbor interactions [149]. This model describes a system of spins that can be in q different states. The interaction is ferromagnetic, i.e. it favors spin alignment, so at zero temperature all spins are in the same state. In this work spin variables are assigned to the vertices of a graph with community structure, and the interactions are between neighboring spins.

Finally, Table 2.1 shows a summary of the main partitional methods for CD that have been previously mentioned, providing a brief description for each one.

2.2.3 Overlapping Methods

A partition is a division of a graph into clusters, or communities, where each vertex belongs to only one cluster. But in real networks vertices may belong to different communities. In these cases, it would be interesting to perform a division of the graph into overlapping communities. In Xie et al. [178] a review of the state of the art in overlapping community detection algorithms is shown. A common feature observed by several algorithms in real-world networks, is the relatively small fraction of overlapped nodes (typically less than 30%), usually each of them belongs only to 2 or 3 communities. Therefore, the results of this work suggest that the overlapping community detection in real networks is still not yet fully solved.

Generally, there are two types of assignments in these type of algorithms: crisp (non-fuzzy) or fuzzy assignment. In a crisp assignment, the relationship between a node and a cluster is binary, whereas in a fuzzy assignment, each node is associated to a community with a belonging factor. Fuzzy community detection algorithms quantify the strength of association between all pairs of nodes and communities. In these algorithms, a belonging factor [87] is calculated for each node. A method combining spectral mapping, fuzzy clustering and the optimization of a quality function, has been presented by Zhang et al. [186]. The algorithm consists on three phases: Firstly, vertices are embedded in an Euclidean space; Secondly, the corresponding vertex points are grouped in a given number of clusters; And finally, a modularity function is maximized over the set of groups found in the second step.

Name	Author	Type	Description
EBC	Girvan and Newman [133]	Divisive	Iteratively removes links with the highest betweenness score
Clauset-Newman-Moore	Clauset et al. [41]	Modularity Optimization	Using matrix of modularity variations
Fastgreedy	Newman [132]	Modularity Optimization	Improves the high computational complexity based on the greedy optimization of modularity
Leading Eigenvector	Newman [131]	Modularity Optimization	Reformulate the modularity measure in terms of eigenvectors
Multi-Level	Blondel et al. [28]	Modularity Optimization	Iterative algorithm that calculates modularity based on nearest neighbours
NetWalk	Zhou and Lipowsky [187]	Random Walks	Agglomerative hierarchical method where the similarity between vertices is expressed by their proximity
Walktrap	Pons et al. [146]	Random Walks	Agglomerative hierarchical method based on the idea that short random walks tend to stay in the same community.
Infomap	Rosvall et al. [152]	Random Walks	Uses the probability flow of random walks on a network to generate regularity maps of the structure
Label Propagation	Raghavan et al. [147].	Iterative Method	All the vertices are iteratively labeled with the dominant label of its neighbourhood
Reichardt-Bornholdt	Reichardt and Bornholdt [149]	Statistical Model	Mapping the graph onto a zero-temperature q-Potts model with nearest-neighbour interactions

Table 2.1: Summary of Partitional methods for CDP.

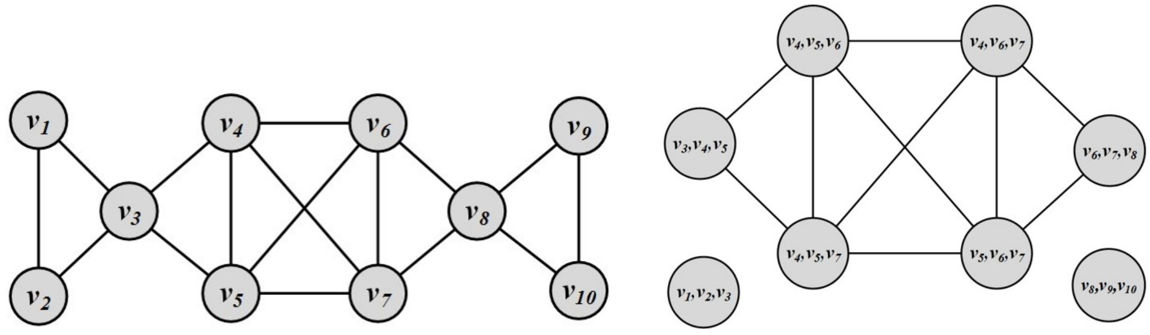
One of the most popular non-fuzzy overlapping technique is the ***Clique Percolation Method (CPM)***. This method tries to find communities using k-cliques (this concept was described in section 2.1.1) [138]. As shown in Algorithm 2, this method starts computing all the cliques of size k into the graph. Then a clique graph is created where all the cliques are represented as nodes, and if two cliques share $k-1$ nodes is represented as an edge between them. Finally, the outcome of the algorithm will be the connected components of this clique graph which represent the found communities.

Figure 2.10 shows the corresponding clique graph generated by the CPM algorithm for the graph shown in the subsection (a) of this figure. In the example the cliques of sizes equal to 3 ($k = 3$) have been computed to create the clique graph. As described in Algorithm 2, the

Algorithm 2: Clique Percolation Method

```
Input: A graph $G = (V, E)$ where $V$ is a set of vertices denoted by $\{v_1, \dots, v_n\}$ and $E$ is
       a set of edges $E$ denoted by $e_{ij}$ representing whether there is a connection
       between the vertices $v_i$ and $v_j$. And a positive number $k$ representing the clique
       size.
Output: Overlapping Communities extracted from connected components of
        cliqueGraph (CG)
$Cliques \leftarrow calculateCliquesSize(G, k)$
$VC \leftarrow \emptyset$
$EC \leftarrow \emptyset$
for $i \leftarrow 1$ to $|Cliques|$ do
    for $j \leftarrow 1$ to $|Cliques|$ do
        $VC \leftarrow VC + vClique_i$
        $VC \leftarrow VC + vClique_j$
        $nShareNodes = calculateShareNodes(clique_i, clique_j)$
        if $nShareNodes > k - 1 \wedge i \neq j$ then
            $EC \leftarrow EC + eClique_{ij}$
return $ConnectedComponents(CG)$ where $CG = (VC, EC)$
```

connected components of the clique graph are returned as communities. In this example, the CPM resulting communities will be: $\{1, 2, 3\}\{8, 9, 10\}\{3, 4, 5, 6, 7, 8\}$. It can be noticed that nodes 3 and 8 belong to two communities, so these communities are overlapping.



(a) Original Graph **(b)** Clique Graph

Figure 2.10: Example of Clique Graph (b) created by CPM algorithm using the cliques ($k = 3$) obtained from a graph (a). This example has been taken from [184].

The complexity of CPM procedure can be very high, and the computational time needed to find all k-cliques of a graph is an exponentially growing function related to the graph size. Therefore, Kumpula et al. [107] have developed a fast implementation of the CPM, named the Sequential Clique Percolation algorithm (SCP). It works detecting k-clique communities by sequentially inserting the edges of the graph at study, one by one, starting from an initial empty graph.

The Cluster-Overlapping Newman Girvan Algorithm (***CONGA***) [85] extends EBC divisive algorithm allowing to a node be a member of multiple communities. In EBC the basic operation is removing an edge, and this extended version introduces a second operation to split vertices. To provide a way to decide when a node should be split, a new concept named split betweenness is introduced. In this work, the split betweenness of a vertex v is defined as the number of shortest paths that would pass between the two parts of v if it were split. The EBC algorithm is extend so in each step is considered the split betweenness of every vertex as well as the edge betweenness. And, if the maximum split betweenness is greater than the maximum edge betweenness, the corresponding vertex is split. However, this algorithm inherits the high computational complexity of EBC algorithm.

For this reason, a new version named CONGA Optimized Algorithm (***CONGO***) [86] was implemented using a local betweenness measure to optimize the computational time. The edge betweenness is expensive to compute because it counts all shortest paths in the graph. One way to avoid this is to count only maximum number of the shortest paths. Therefore, for CONGO algorithm, is used the local edge betweenness that is defined as the number of shortest paths running along the edge whose length is less than or equal to h (a parameter of the algorithm).

A model-based approach was proposed by Yang et al. [181] that considers an underlying model of statistical nature to generate the division of the network. The Cluster Affilitation Model for Big Networks algorithm (***BIGCLAM***) was designed for networks where the overlaps of communities are more densely connected than the non-overlapping parts. This algorithm is based on the notion that nodes in the same community are more likely to share an edge. It captures the probability that a pair of nodes are connected as a function of that shared membership, and formulates community detection as a variant of Non-negative Matrix Factorization (NMF), which aims to learn factors that can recover the adjacency matrix of a given network.

Subsequently, Communities through Directed Affiliations algorithm (***CoDA***) appears as an extension of the previous algorithm incorporating the use of edges to discover communities [183]. This improvement allows CoDA algorithm to hanlde networks with millions of nodes. In addition, its nature allows to detect communities in directed as well as undirected networks. It extends the traditional definitions of network communities including the concept of dense interlinked nodes in two different ways: in cohesive communities (a node can sends and receives links from other members) or 2-mode communities (modelled with unidirectional memberships where some members mostly send/create links while others mostly receive from them). Other model-based approach was proposed by Yang et al [181] that considers an underlying model of statistical nature that can generate the division of the network.

Other approach to discover community structure in a graph is based on the idea of partitioning links instead of nodes. Recent studies suggested that defining clusters as sets of edges can be a promising strategy to analyse graphs with overlapping communities [64]. Ahn et al. [8] proposed to group edges with an agglomerative hierarchical clustering technique, called hierarchical link clustering. They use a similarity measure for a pair of (adjacent) edges that expresses the size of the overlap between the neighbourhoods. Groups of edges are merged pairwise in descending order of similarity, until all the edges are together in the same cluster. In Kim and Jeong [105] work, the Infomap algorithm is extended to the line graph, which encodes the path of the random walk on the line network under the Minimum Description Length (MDL) principle.

Finally, Table 2.2 presents a summary of the main overlapping methods for CD that were previously introduced, providing a brief description of them.

Name	Author	Type	Description
CPM	Palla et al. [138]	Clique Discovery	Find communities using k-cliques where k is an input parameter
CONGA	Gregory [85]	Divisive	Extended version of EBC with second operation to split vertices based on a new concept (split betweenness)
CONGO	Gregory [86]	Divisive	Extended version of CONGA using a local betweenness that is defined as the number of shortest paths running along the edge
MakeFuzzy	Gregory [87]	Modularity Optimization	Based on a belonging probability factor per node
BIGCLAM	Yang et al. [181]	Model Based	Formulate the CDP as a variant of NMF using a probability that a pair of nodes are connected as a function of that shared membership
CoDA	Yang et al. [183]	Model Based	Extension of BIGCLAM that incorporates the use of edges to discover communities
Evans-Lambiotte	Evans and Lambiotte [64]	Edge Divisive	Defining clusters as sets of edges instead nodes
Hierarchical link clustering	Ahn et al. [8]	Edge Divisive	Use a similarity measure for a pair of (adjacent) edges that expresses the size of the overlap between the neighbourhoods
Kim-Jeong	Kim and Jeong [105]	Random Walks	Infomap algorithm is extended to encode the path of the random walk on the line network under the MDL principle
Zhang	Zhang et al. [186]	Fuzzy clustering and Quality Optimization	Combining spectral mapping, fuzzy c-means clustering and the optimization of a quality function

Table 2.2: Summary of Overlapping methods for CDP.

2.2.4 Evolutionary Approaches

This section provides a briefly introduction to the different types of algorithms existing in the Evolutionary Computation area have been introduced, and how these algorithms have been applied to the CDP.

2.2.4.1 Introduction on Evolutionary Algorithms

Evolutionary Computation (EC) algorithms are a population-based algorithms that search through the space of possible solutions using the Darwin's evolution theory based on the natural selection and the survival of the fittest.

All EC algorithms share three main characteristics [14, 60]: (1) they use a **population** of **individuals** that represent solutions, (2) individuals are modified using at least one **genetic operator**, and (3) individuals are under a selective pressure, where the fittest ones can pass their characteristics to the next generation. From an AI point of view, EC is a set of stochastic search algorithms.

As previously mentioned, each individual represents a possible solution to the problem, and its characteristics are encoded into different *genes* that compose the *genotype*. Therefore, the genotype describes the genetic composition of an individual and the genes represent the smallest, and inheritable, units of information. Also, individuals contain a *phenotype*, that is the representation of the genotype in the solution space.

In EC the fittest individuals are determined by the *fitness function* that provides the quality of a given individual of the population. The fitness value of any individual will determine its reproduction probability, i.e. individuals with better fitness will have more probabilities to reproduce offspring for the next generation and thus transmit its genes.

The general process of any EC algorithm has been shown in Algorithm 3. The algorithm is composed by a population (P_t) that contains n randomly initialized individuals. Once the population has been initialized, the evolution process starts. Initially, every individual in the population is evaluated (line 4) using the fitness function. Then, one, or more, parents will be selected (line 7) from the population where usually individuals with "better" fitness will have more probabilities for being selected. Once the parents have been selected, the new offspring are generated by the reproduction process (line 8). Finally, when the new population is fully generated, the process starts again evaluating the new population.

Algorithm 3: Generic Evolutionary Algorithm

```
1  $t \leftarrow 0$
2  $P_t \leftarrow n$ initialized individuals
3  while termination criteria is not satisfied do
4      EvaluateFitness($I_i$) $\forall I_i \in P_t$
5      $P^* \leftarrow$ new empty population
6      while $P^*$ is not full do
7          parents $\leftarrow$ selectParent($P_t$)
8          offspring $\leftarrow$ reproductionProcess(parents)
9          includeOffspring($P^*$,offspring)
10     end
11     $t \leftarrow t + 1$
12     $P_t \leftarrow P^*$
13 end
```

This process is repeated until a convergence criteria is satisfied, or a budget of computational resources (time, iterations, etc) is exhausted. The complexity of the algorithm depends on the encoding, and the operators that are used to perform the complete evolutive process [44, 165]. The most common genetic operators int the reproduction process are the **mutation** and **crossover**. The first one introduces random changes in the individuals, while the second ones imitates the reproduction process, joining the genetic material of two individuals to generate a new offspring. The operators of the genetic algorithms can also be modified depending on the application domain. Some examples of these modifications can be found in Poli and Langdon approach [144] where the algorithm is improved through backward-chaining, creating and evaluating individuals recursively reducing the computation time.

When a EC algorithm is applied to any problem, there are three issues that must be taken into account:

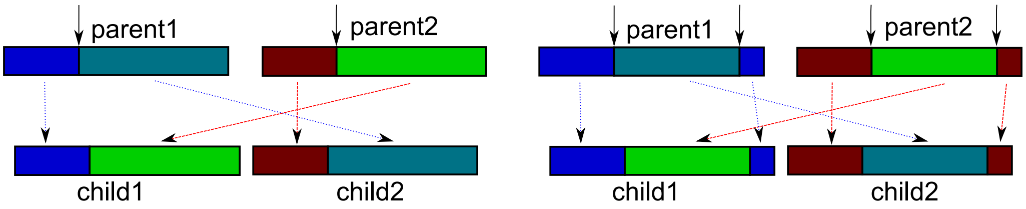
1. **Genotype encoding**. The encoding of the genotype influences the selection and reproduction processes. Also, the encoding influences in the performance of the algorithm.

2. **Selection operator**. This operator is used in order to select those individuals that will produce the offspring for the next generation. The most commonly used selection operators are: *random*, *proportional*, and *tournament* selection. The first one is the simplest selection operator where all the individuals have the same probability to be selected. In the second operator, better individuals have more probabilities to be selected. A popular sampling method, that uses the proportional operator, is called *roulette wheel sampling.* Assuming that the fitness values are normalized and the goal is to maximize the fitness function, The probability distribution can be seen as a roulette wheel, where the size of each slice is proportional to the normalized selection probability computed. Finally, in the tournament selection operator, individuals are randomly selected from the population. Then the fitness values for the chosen individuals are compared and the individual with the best fitness is selected.

3. **Reproduction process**. The reproduction process is the responsible of generating the new offspring once the parents have been selected. This process is composed by two operators, *crossover* and *mutation.* The **crossover** operator generates a new offspring by combining the genetic material of two, or more, parents which have been selected using one of the methods described in the selection operator. There are different crossover techniques in the literature, but the simplest crossover operators are called *One-Point* and *Two-Points* crossover [63, 90] (see Figure 2.11). Finally, the **mutation** operator is the process of randomly changing the values of the genes in a chromosome. The main goal of this operator is to increase the genetic diversity. If the genes contains binary data, the application of this operator consists in selecting the genes randomly, and according to a mutation probability, the corresponding bit values are negated. In the case that the genotype contains more complex genes, the mutation operator consists in including a variation to the value of these genes.

Depending on how individuals are represented and how they are modified, several types of algorithms can be found. The four fundamental families of this kind of algorithms are listed below:

- **Genetic Algorithms** (GAs): These algorithms are based on a linear vector representation (usually binary or integer) in the genotype [82, 90], and they are commonly used as function



(a) One-Point Crossover **(b)** Two-Points Crossover

Figure 2.11: Graphical representation of one-point, and two-points crossover for a GA.

optimization methods. The individual selection is carried out according to their fitness function values, where the better individuals will be selected with a greater probability. On the other hand, the mutation probability is usually low, so the solution diversity on the search in these algorithms are based on the recombination of the individuals to generate new solutions. This is the most distinctive feature of this type of algorithms.

- **Genetic Programming** (GP): Its main feature is that the individuals of the population are programs represented by trees [106, 145]. In this case, the initialization mechanisms and variation operators have to deal with the problem of generating non-valid individuals. The crossover operator exchanges random sub-trees of two parents, while the mutation operator introduces a new sub-tree into a random node of the tree. Finally, the evaluation of individuals is as simple as executing the program that represents each one, and analysing it result to determine how it works.

- **Evolutionary Programming** (EP): where the structure that is evolved is based on a finite automata [70, 71]. The distinctive feature of these algorithms is the emphasis that make on the relationship between the parents and their offspring, rather than the genetic relationship which is taken into account in GAs. Each individual of the population is a progenitor, generating λ offspring from it using only the mutation operator. In this family of EC algorithms the crossover operator is not used.

- **Evolutionary Strategies** (ES): They are focused on the numeral optimization in the space of real numbers which are normally used in the optimization of problems with continuous parameters [24, 158]. Generally, the mutation applied to individuals in this algorithm is a random change following a normal distribution $N(\mu, \sigma)$. On the other hand, the crossover usually consists on the random selection of an element, from different individuals, swapping its values. The selection of these individuals is deterministic according to the order that each individual occupies in the ranking of their fitness function values, and it is not biased by these values. Regarding to the encoding of individuals, a vector of real numbers is usually used where the parameters μ and σ defined for the mutation are included. This lets their co-evolution at the same time as the individuals, giving rise to a new methodology that has been termed self-adaptation. The self-adaptation is the most important contribution of the ES to the field of evolutionary computation, and this improvement has also later use in other evolutionary algorithm families.

All the techniques previously mentioned about EC come from the premise that only has one objective to be optimized as fitness function. However, many of the real-world problems are *multi-objective* problems, where the quality of a given solution is defined by its performance according to several different objectives. A possible approach to deal these problems is to assign a numerical weight for each objective according to it importance to measure the quality of the solution. Finally, all these different objectives are grouped into a single value by a weighted sum, known as scalarization [42].

Other option is to consider jointly all the objectives and establish a single quality value among them. For this purpose is necessary to define the concept of *dominant solution*. One solution is dominant with respect to another solution, if it improves the other solution for all objectives. On the contrary, a solution is called *non-dominated*, or *Pareto optimal*, with respect to a set of objectives, if there exists no other solution in the decision space, that improves simultaneously all its objectives. All Pareto optimal solutions are considered equally good, because them simultaneously optimize each objective. Therefore, there are several multi-objective evolutionary algorithms based on the finding of the representative set of Pareto optimal solutions, such as *NSGA-II* [53] or the *SPEA-2* [188].

2.2.4.2 Applications of GAs to Graph Clustering

GAs [90] have been traditionally used in a large number of different domains such as Data Mining [18, 125], Mission Planning [148], Malware detection a classification [121] or optimization problems [26, 39, 51, 65, 75, 155], amongst others. GAs have been demonstrated to be robust algorithms able to find satisfactory solutions in highly multidimensional problems with complex relationships between the variables. In this kind of problems, GAs are able to find a optimal solutions in a large search space evaluating a minimum number of points.

The evolutionary approaches for *clustering* try to improve the results using different fitness functions to tune up the cluster sets selection. In these approaches, the encoding and optimization algorithm are used to look for the best set of groups optimizing a particular feature of the data. Hruschka et al. [93] provide a complete review on evolutionary algorithms for clustering. There are several methods based on evolutionary approaches from different perspectives, for example: modifing the fitness considering cluster asymmetry, coverage and specific information of the studied case [7]; using a compact spherical cluster structure and a heuristic strategy to find the optimal number of clusters [168]; using the clustering algorithm for metric optimization trying to improve the cluster centre positions [122]; applyng an Extend Classifier Systems that is a kind of Learning Classifier System to guide the search, in which a fitness of the classifier is determined by the measure of its prediction's accuracy [164]; using Differential Evolution that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality [51].

In particular to solve *graph clustering* problems, GAs have been used to look for the best partition of the graph in groups of nodes (clusters) optimizing a particular feature of the graph structure. Several graph clustering algorithms such as K-means or Fuzzy C-means [22, 116, 123, 127] have been improved using GAs techniques.

Regarding the optimization criteria used to detect the clusters (group of nodes), there are serveral approaches in the literature. Many evolutionary graph clustering algorithms use a single optimization criteria as the objective function, being the modularity one of the most common criteria used [161]. However, there are also works where the community detection is solved as a multi-objective optimization problem. In Gong et al. [83] a multi-objective evolutionary algorithm based on graph decomposition is proposed maximizing the density of internal degrees, and minimizing the density of external degrees simultaneously. The multi-objective genetic algorithm for networks (MOGA-Net) [143] optimizes two objective functions to identify densely connected groups of nodes having sparse inter-connections. The first objective function uses the concept of community score [109] to measure the quality of the network division into communities, whereas the second defines the concept of node fitness [141] measuring the membership of a node to a module. Then the algorithm iteratively finds modules having the highest sum of node fitness. Finally, Liu et al. [118] design three objective functions to guide a multi-objective evolutionary algorithm measuring quality, separation and overlapping communities respectively.

Other different approach was proposed by Pizzuti [142], where a GA is used to detect overlapping communities using a line graph as input the algorithm (GA-Net+). In the line graph, nodes represent the edges of the original node graph, while edges represent the adjacent relationships of original edges. So, this method requires that for each iteration, the communities in the line graph are converted into communities of the original graph. This transformation between line graph and node graph has a highly computationally cost, so this fact decreases the effectiveness of this method.

All these algorithms are based on the idea of node clustering. However, as mentioned in the description of overlapping methods, recently studies have suggested that partitioning the graph into clusters of edges is a good strategy to identify overlapping communities. For this purpose, Shi et. al. [163] proposed a GA for edge clustering of graphs (GaoCD) by optimizing the Density as objective function. The most significant contribution of this work is a new encoding schema based on the edges, where the number of communities can be automatically determined. The experimental results of this work show that GaoCD algorithm can effectively identify overlapping communities, and moreover it can be able to detect larger communities which is more useful for real world networks. This is due to real world networks are usually very sparse, and the cliques contained in them are often small, so the CD algorithms tend to find small communities.

Finally, as in the previous sections, Table 2.2 shows a summary of the main approaches based on GAs for CD that have been presented in this section. This table provides a brief description for each one.

2.3 Applications

The Web is one of the most important sources of data in the world producing every day vast amounts of public information. The exponentially increasing of websites and online web services in the last years has allowed new interdisciplinary challenges for several fields and computer science, such as Marketing Campaigns [17, 23, 38], Financial Prediction [13] or Public Healthcare [37, 45, 97], amongst others.

The next subsections introduce two application examples of data extracted from the Web and SN to acquire new knowledge by applying data mining techniques. Both examples will be later used as application domains to test the new algorithms proposed in this thesis.

Name	Author	Type	Description
Shang	Shang et al. [161]	GA	Modularity as single optimization criteria
Gong	Gong et al. [83]	MOGA	Maximizing the density of internal degrees, and minimizing the density of external degrees simultaneously
GA-Net+	Pizzuti [142]	GA	Using a line graph in the GA to detect communities instead of the original graph
MOGA-Net	Pizzuti [143]	MOGA	Using as objective functions a score to measure the quality of the network division into communities, and a measure of the membership of a node to a module
Liu	Liu et al. [118]	MOGA	Three objective functions measuring quality, separation and overlapping communities respectively
GaoCD	Shi et. al. [163]	GA	With edge-based encoding, and optimizing the density as objective function

Table 2.3: Summary of Evolutionary approaches for CDP.

2.3.1 Eurovision Song Contest

The Eurovision Song Contest can be understood as a complex system [29], where interactions between countries are heavily influenced by factors like geography, shared history, culture and migration patterns. Voting patterns for each country seem to be dictated, not by the artistic value of the song, but by a latent affinity between countries. These voting patterns revalue the contest as more than a search for current trends in music, because it provides an active forum, where countries are free to give opinions about the rest of the participants without fear of economic or political backlash [66, 140].

2.3.1.1 Historical Background

The Eurovision song contest is an annual competition among members of the European Broadcasting Union [59], running continuously ever since it's inauguration in 1956. In this contest, each country submits a song and a performer to compete. All songs are then performed live, in a transmission available to all participating countries. Once all songs have been performed, votes are casted and a winner is selected.

The contest has undergone a series of changes throughout the years, in an effort to keep it fresh and maximize viewer attention. From 1956 to 1996, votes where cast by a jury of representatives sent from each of participating countries. Jurors then cast all of ten individual point-votes ranging from 1-8, 10 and 12 points with no repetitions. In 1997 *televoting* was introduced in five countries, to gradually displace the jury-based system until 2004 when *televoting* was made mandatory for all participants.

Televote technology allows viewers to cast their votes via phone, sms or the internet for a set window of time, normally within the live broadcast. After the voting window closes, all votes are tallied, and points are given in decreasing order: the participant with most votes receives 12 points, the next highest receives 10, and so on.

In 2004 a *semi-finals* round was introduced to offset the increasing number of participants in the contest. In order to participate in the Eurovision contest, participants must clear this preliminary round, thereby limiting the number of participants to a manageable size. That year's host-country and the so-called "Big Four" are exempt from this filter. The "Big Four" being the four highest contest contributors: France, Germany, Spain and the United Kingdom. However, all countries, finalists or not, are allowed to vote in the final round, which inflates the number of countries that vote and overall score of the winners each year. Critics contested that because of migration patterns, *televoting* had a tendency to favour certain countries, 2009 saw the implementation of the current voting system. A hybrid system of *televoting* and a jury was implemented, whereby each part contributes half of the total vote tally for each country.

2.3.1.2 Studies on the Eurovision Contest

The Eurovision contest has been analyzed by applying different data mining techniques such as clustering methods [179, 180], regression analysis [79], dynamical networks [66], or analytical identification of statistically significant tends [77]. However, all the previous research works were focused on grouping the participant countries into blocks of similar behaviour, acquiring new social knowledge about European Union.

One of the earliest analyses was carried out by Yair [179, 180], which presents an analysis based on the annual voting matrix of the contest, using network analysis to identify the generic structure of the relationships between the participant countries. The study contemplated two perspectives, firstly looking for existing clusters or cliques into the network. And secondly, detecting those members who have an equivalent role within the network, and therefore a similar voting pattern. The results related to the clustering detection revealed that the Eurovision community was split into three main blocs (see Table 2.4): The *Mediterranean Bloc*, the *North Bloc* and the *Western Bloc*. In this model, the western bloc consistently amassed the highest number of votes, and was the largest of the three. Briefly, the Western block can be seen as a coalition based on historical and political interests. The North Block have common languages (German roots), and primarily cultural similarities. And finally, the Mediterranean block is the most unstable alliance based on their similar cultures. In contrast to the blocks found with the first method, the second analysis obtained a greater dispersion identifying blocks much more varied. These groups mostly have similarities in language and cultural roots. However, it is very interesting that again the Western block has the most important role.

Block Name	Belonging Countries
Mediterranean	Italy, Greece, Spain, Yugoslavia, Turkey and Monaco
North	Germany, Sweden, Norway, Denmark and Belgium
Western	England, Ireland, France, Netherlands, Switzerland, Malta, Luxembourg and Israel

Table 2.4: Blocks of participant countries in Eurovision detected by Yair work [180].

In a first work carried out by Derek Gatherer [76], it was analysed the period from 1975 to 2002 using statistical techniques, and two large blocks were identified : *The Viking Empire* (Scandinavian and Baltic countries) and *The Warsaw Pact* (Russia, Romania and the old republic of Yugoslavia). In a further paper, this author presented a non-analytical approach to the problem, simulating each year according to a Monte Carlo method. This allows the empirical estimation of an statistical threshold to compare the real data against the simulated data. When a country votes above the threshold to another country (reciprocally) is identified as a vote pattern. The results showed that the two large blocks detected in his previous work were growing, being more influential from 2001. In 2003 and 2005 the vote of the "Balkan Block" was able to make a member belonging to it win, as happened in 1999 and 2002 with the "Viking Empire" block. Moreover, these results demonstrate the emergence of strong voting patterns. If only the contest result depended on the quality of the songs, comparing with a enough number of years the distribution of votes should be random. But, this fact didn't happened in this study.

On the other hand, Fenn et al. [66] used dynamic network analysis to study voting partnerships, observing that these may not be static, but are instead susceptible to change over time. This paper performed a network analysis on two levels: an initial study of the *global* properties of the network; and later an analysis focused on *local* properties of specific parts (nodes and edges). In the global analysis, the clustering coefficient for each year were calculated and compared against the clustering coefficient values obtained from a random network. The clustering coefficient value was always greater than the value obtained from a random graph, therefore this demonstrated the existence of voting patterns. Then, the countries were grouped according to a similarly behaviour using clustering techniques, and to measure the distance between countries, it was used the average of votes cast by those countries (see Figure 2.12). Finally, analysing from a local perspective the links between the nodes belonging to each cluster, UK had the greater number of reciprocal and stable links, contradicting the popular notion that UK is often not consistent with the European Union. On the opposite side, countries such as France or Spain had the smallest number of stable links.

Finally, Ginsburgh and Noury [79] analysed data from 1975 to 2003 using a regression equation to compare the votings of a given year with the previous two years. This equation contained variables on linguistic and cultural distances between countries. The authors found voting patterns that disappeared by applying a correction of linguistic and cultural distances. This fact confirmed that countries were voting according to pre-established cultural trends rather than based on the alliance formation (for example, the vote of immigrants).

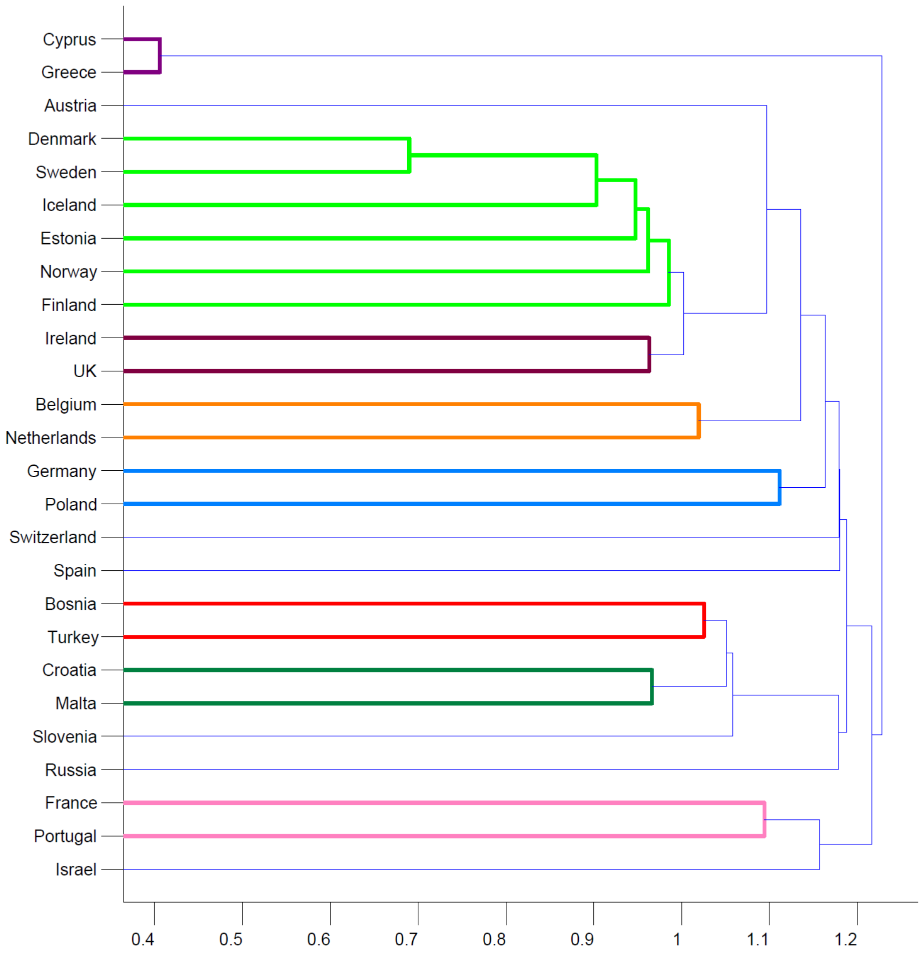


Figure 2.12: Dendrogram showing the voting clusters within Eurovision graph taken from [66].

2.3.2 Public Healthcare

The research on AI techniques, applied to develop technologies allowing the monitoring of web data sources for detecting public health events, has been emerged as a new relevant discipline called Epidemic Intelligence (EI) [46, 88, 97, 139].

EI can be defined as the early identification, assessment, and verification of potential public health risks [139], and the timely dissemination of the appropriate alerts. This discipline includes surveillance techniques such as automated and continuous analysis of unstructured free text information available on Web from social networks, blogs, digital news media or official sources. Surveillance systems are nowadays widely used by public health organizations such as World Health Organization (WHO) or the European Centre for Disease Prevention and Control (ECDC) [88]. Tracking and monitoring mechanisms for early detection are critical in reducing the impact of epidemics giving a rapid response. For instance, several of these systems are able to discover early events of the disease breakout during the A(H1N1) influenza pandemic in 2009 [46].

Traditional epidemic surveillance systems are implemented using virology and clinical data, which is manually collected, and these traditional systems often have a delay reporting the emerging diseases. But in situations like epidemic outbreaks, real-time feedback and a rapid response is critical. Social Web media is a profitable medium to extract the society opinion in real time. Blogs, micro-blogs (Twitter), and social networks (Facebook), enable people to publish their personal opinions in real time, including geo-information about their current locations. These big data, including situation and context aware information about the users, provide an useful source for public healthcare. However, the extraction of information from the web is a difficult task due to its unstructured definition, high heterogeneity, and dynamically changing nature. Because of this diversity in the data format, several computational methods are required for its processing and analysing (Data Mining, Natural Language Processes (NLP), knowledge extraction, context awareness, etc...) [97].

This section presents a survey on current data mining applications, and web systems, based on web data for public healthcare over the last years. It tries to take special attention to machine learning and data mining techniques, and how they have been applied to these web data sources, to extract collective knowledge from social networks as Twitter.

2.3.2.1 Epidemic Intelligence Systems

Nowadays, large amounts of emergency and health data are increasingly coming from a large range of web and social media sources. This information can be very useful for disease surveillance and early outbreak detection, and several public surveillance projects in this field have emerged over the recent years.

One of the earliest surveillance systems is the Global Public Health Intelligence Network (GPHIN) [126] developed by Public Health Agency of Canada in collaboration with WHO. It is a secure web-based multilingual warning tool, which is continuously monitoring and analysing global media data sources, to identify information about disease outbreaks and other events related to public healthcare. The information is filtered for relevancy by an automated process, and categorized based on a specific taxonomy of categories. From 2002 to 2003 years, this surveillance system was able to detect the outbreak of Severe Acute Respiratory Syndrome (SARS).

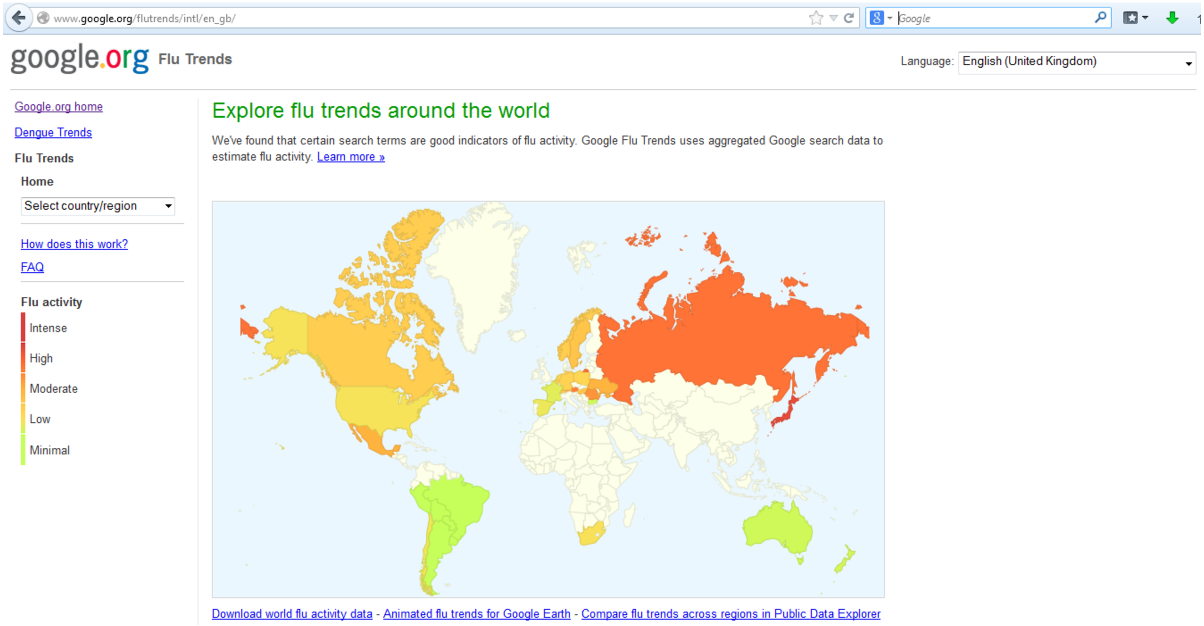
Since 2006, BioCaster [46] is an operational ontology-based system for monitoring online media data. This system is based on text mining techniques for detecting and tracking the infectious disease outbreaks through the search of linguistic signals. The system continuously analyses documents reported from over 1700 RSS feeds, Google News, WHO, ProMED-mail, and the European Media Monitor, among others providers. The extracted text are classified by their relevance, and plots them onto a Google map using geo-information. The system consists on four main stages: topic classification, named-entity recognition (NER), disease/location detection, and event recognition. In the first stage, the texts are classified into relevant or non-relevant categories, which are later used to train a naive Bayes classifier.

HealthMap project [35] is a global disease alert map that uses data from different sources such as Google News, expert-curated discussion such as ProMED-mail, and official organization reports such as World Health Organization (WHO) or Euro Surveillance. This is an automated real-time system that monitors, organizes, integrates, filters, visualizes, and disseminates online information related to emerging diseases.

Other system that collects news from the Web, related to human and animal health, is EpiSpider [103], plotting the data on a Google Maps mashupare. This tool automatically extracts infectious disease outbreak information from several sources, including ProMed-mail and medical web sites, and it is used as a surveillance system by public healthcare organizations, several universities, and health research organizations. In addition, this system automatically converts these topics and location information of the reports into RSS feeds.

MedISys system [117] is also used by a Public Health Organization (The European Centre of Disease Prevention and Control), monitoring human and animal infectious diseases, as well as chemical, biological, radiological and nuclear (CBRN) threats in open-source media. This system automatically collects articles concerning public health in various languages from news, which are classified according to pre-defined categories.Users can display world maps in which event locations are highlighted as well as statistics on the reporting about diseases, countries and combinations of them.

Google search data is used by Google Flu Trends [36] to estimate flu activity during two weeks giving an early detection of disease activity, as shown in Figure 2.13. This web service correlates search term frequency with influenza statistics reported by the Centers for Disease Control and Prevention (CDC). It allows a quicker response in a potential pandemic of influenza, thus reducing its impact. Internet users perform search queries [78], and post entries in blogs using terms related to influenza illness as its diagnosis and symptoms. An increase, or decrease, in the number of illness searches and posts in blogs, reflects a higher or lower potential outbreak focus for influenza illness.



www.google.org/flutrends/intl/en_gb/

Language: English (United Kingdom)

Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

Select country/region

Figure 2.13: Google Flu Trends showing flu activity during two weeks. This screenshot has been taken from the Google Flu Trends Web (http://www.google.org/flutrends/).

Finally all previous mentioned systems, with their main characteristics, have been listed in Table 2.5.

Name	Website	Data Sources	Description
GPHIN	-	News wires and Web Sites	Warning tool to detect disease outbreaks
BioCaster	http://born.nii.ac.jp	RSS feeds, Google News, WHO, ProMED-mail and European Media Monitor	Ontology-based system for monitoring online media data
HealthMap	http://www.healthmap.org	Google News, ProMED-mail, WHO and Euro Surveillance	Global disease alert map
EpiSpider	http://www.epispider.org/	ProMed-mail and medical web sites	Human and animal disease alert map
MedISys	http://medusa.jrc.it/ medisys/homeedition/ en/home.html	Articles concerning public health from news	Monitoring tool for human and animal infectious diseases and chemical, biological, radiological and nuclear threats
Google Flu Trends	http://www.google.org/ flutrends/	Google search and CDC reports	Monitoring system of influenza

Table 2.5: Summary of Epidemic Intelligence Systems.

2.3.2.2 Disease outbreaks Detection

Text mining techniques have been widely applied to biomedical text corpus for NER, text classification, terminology extraction, or relationship extraction [43]. These methods are human language processing algorithms that allow to convert unstructured textual data, from large-scale collections into a specific format, filtering them according to the needs of the application domain.

The text analysis methods can be used for trend detection, identifying potential sources of disease outbreaks. But this goal can be difficult because the same word can refer to a different thing depending upon context. Furthermore, a specific disease can have multiple names and symptoms associated, which increases the complexity of the problem. Ontologies can help to automate human understanding of key concepts and relations between them and allow that a better level of filtering accuracy can be achieved. Biomedical ontologies contain lists of terms and their human definitions, which are then given as unique identifiers and classified into classes with common properties according to the specific domain treated. Currently there are various available ontologies that contain all the biomedical terms necessary. For example, BioCaster ontology (BCO) [47] is in the OWL Semantic Web language to support automated reasoning across terms in 12 languages.

Discovering the time and location of the text is a crucial value added by EI systems to be able to detect disease outbreaks as quickly as possible, and thus to take the best decisions in order to control them. In practice, location names are often highly ambiguous because geo-temporal disambiguation is so difficult, and because of the variety of ways in which cases are described across different texts. Keller et al. [104] work provides a review of the issues for epidemic surveillance, and present a new method for tackling the identification of a disease outbreak location based on neural networks, which are trained on surface feature patterns in a window around geo-entity expressions.

A new unsupervised machine learning approach is proposed by Fisichella et al. [69] to detect public health events. This new approach defines a generative model for predictive event detection from a document by modelling the features based on trajectory distributions. The novelty of this work is that allows to identify public health events even if cannot be found matching keywords or linguistic patterns.

A different solution for outbreak detection is shown in Leskovec et al. paper [115], where the problem is modelled as a network, in order to detect the spreading of the virus or disease as quickly as possible. They present a new methodology for selecting nodes to detect outbreaks of dynamic processes spreading over a graph. This work shows that many objective functions for detecting outbreaks in networks, such as detection time, likelihood, and population affected, are sub-modular. This means that, for instance, reading only a few blogs provides more new information than reading it after we have read many ones. They use this characteristic to develop an efficient approximation algorithm (CELF) that achieves near-optimal solutions and it is 700 times faster than a simple greedy algorithm.

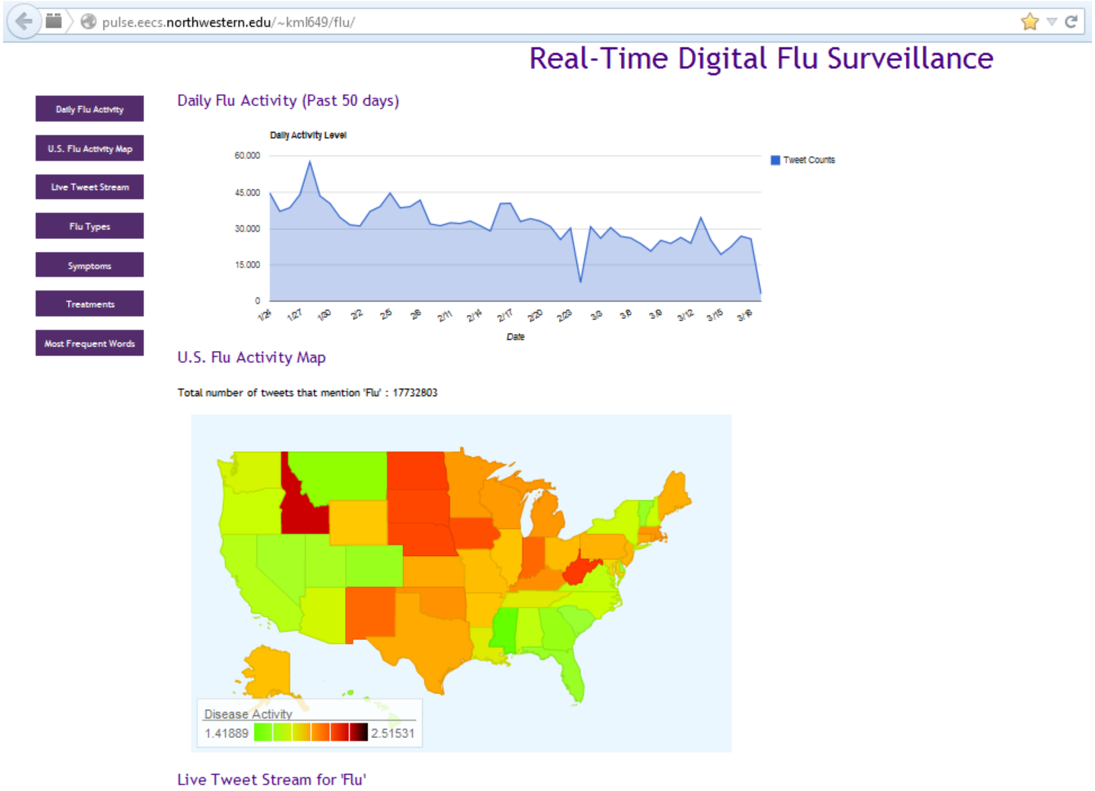
The increasing popularity and use of micro-blogging services, such as Twitter, provides a new valuable data source for web-based surveillance due to its message volume and frequency. Twitter users may post messages related to illness, and in addition their relationships in the network can give information about which people could be in contact with. Furthermore, user posts retrieved from the public Twitter API can come with GPS-based location tags, which can be used to detect potential disease outbreaks for a health surveillance system. Recently, several works have already appeared shown the potential of Twitter messages to track and predict disease outbreaks.

Ritterman et al. [151] work is focused on using prediction market to model public belief about the possibility that H1N1 virus will become a pandemic. In order to forecast the future prices of the prediction market, they decided to use the Support Vector Machine (SVM) algorithm to carry out regression. A document classifier to identify relevant messages is presented in Culotta et al. paper [48]. In this work, Twitter messages related to flu were recollected during 10 weeks using keywords such as flu, cough, sore throat or headache. Then, several classification systems, based on different regression models to correlate these messages with CDC statistics, were compared finding that the best model achieves a correlation value of 0.78 (simple model regression).

Aramaki et al. [11] presents a comparative study of several machine-learning methods to classify tweets related to influenza into two categories: positive or negative. Their experimental results show that SVM model using a polynomial kernel achieves the highest accuracy (FMeasure of 0.756) and the lowest training time.

A novel real-time surveillance system to detect cancer and flu is described in paper [114]. The proposed system continuously extracts text related the two specific diseases from twitter using Twitter streaming API and applies spatial, temporal, and text mining to discover disease-related activities. The output of the three models is summarized as pie charts, time-series graphs, and

US disease activity maps on the project website (http://pulse.eecs.northwestern.edu/ kml649/flu/) as can be seen in Figure 2.14. This system can be useful not only for early prediction of disease outbreaks, but also for monitoring distribution of different cancer types and the effectiveness of the treatments used.



pulse.eecs.northwestern.edu/~kml649/flu/

Real-Time Digital Flu Surveillance

Daily Flu Activity (Past 50 days)

U.S. Flu Activity Map

Total number of tweets that mention 'Flu' : 17732803

Live Tweet Stream for 'Flu'

Figure 2.14: A novel real-time surveillance system to detect cancer and flu from Twitter messages. The screenshot has been taken from the Project Web Site.

Severak well-known regression models are evaluated on their ability to assess disease outbreaks from tweets in Bodnar et al. [30]. Regression methods such as Linear, Multivariable an SVM, are applied to the raw count of tweets that contain at least one of the keywords related to a specific disease, in this case "flu". The results confirmed that even using irrelevant tweets and randomly generated datasets, regression methods were able to assess disease levels comparatively well.

Finally, a summary for all the systems mentioned, and the machine learning techniques used, is listed in Table 2.6. It can be noticed that most of the works use regression models, and these works are usually focused on the detection of influenza outbreaks. These works demonstrate that there are health evidences in social media which can be detected. However, it can appear complications regarding the possible incorrect predictions due to the huge amount of social data existing compared with the small amount of relevant data related to potential diseases outbreaks. Therefore, it is necessary to test and validate carefully all the models and methods used.

Work Name	Machine Learning Techniques	Description
Ritterman et al. [151]	Prediction market model and SVM	Predict flu outbreak detection
Culotta et al [48]	Regression models	Classifier to identify flu relevant messages
Aramaki et al. [11]	SVM using a polynomial kernel	Classifier of influenza tweets into positive or negatives
Lee et al. [114]	Spatial, temporal, and text mining	Surveillance system to detect cancer and flu
Bodnar et al. [30]	Regression models	Disease outbreak detection

Table 2.6: Summary of tracking and monitoring epidemic works using Twitter data.

2.3.2.3 Analysing vaccination sentiments and attitudes

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 [27]. 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 much controversy after the publication of a paper by Andrew Wakefield et al. [170]. 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 [81]. 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. [95] 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 [96]. As shown in Figure 2.15, 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,

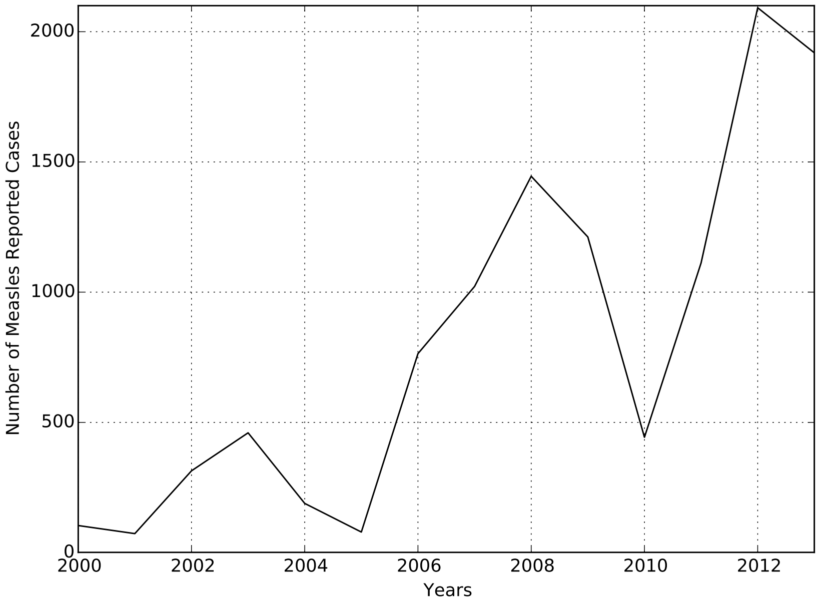


Figure 2.15: Number of measles reported cases between 2000 and 2013 from United Kingdom.

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 changes in virulence or epidemiology, to have vaccines ready, but vaccine supply was insufficient in some areas [112]. 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 leds 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 [100, 111].

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. For this purpose, currently, there are several works related to knowledge acquisition from web sources focused on vaccine sentiments.

VASSA (Vaccine Attitude Surveillance using Semantic Analysis) framework [33] 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, which 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. [32], 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 of relevant 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. [177]. 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. [162] 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 [113]. 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.

Salathe et al. [153] proposed an hybrid approach based on naive Bayes classifier, and a maximum entropy classifier 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. 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.

Finally, a summary for all the approaches mentioned, and the machine learning techniques used, is listed in Table 2.7. It can be noticed that most of the works are focused on the sentimental analysis of the people perceptions about vaccines, and how people can be influenced by social opinions. However, another interesting point of view, it could be to identify and track these groups of persons that are influencing individuals about vaccination, to prevent their possible effects.

Work Name	**Techniques**	**Description**
VASSA [33]	Semantic Web and NLP	Assessment and analysis of vaccination attitudes and beliefs using online data
Botsis et al. [32]	Text mining and several classifiers	Automated text classifier of reports collected from a reporting system about vaccines (VAERS)
Shang et al. [177]	Social impact theory, and game-theoretical analysis	to Framework to simulate and study the social influence on the vaccination decision making
Shaw et al. [162]	Agent-based simulation, mathematical analysis and EAs	Framework to determine the optimal distribution of vaccine supply in a population
Salathe et al. [153]	Naive Bayes and a maximum entropy classifiers	To classify tweets as negative, positive or neutral with respect user intention about influenza H1N1 vaccination

Table 2.7: Summary of works to analysis vaccination sentiments and attitudes.

Chapter 3

Genetic Graph-based Approaches for Overlapping Community Detection

‘Imagination is the discovering faculty, pre-eminently.
It is that which penetrates into the unseen worlds around us,
the worlds of Science.”

- Ada Lovelace

This chapter introduces three soft clustering methods based on a Genetic Algorithms and Graph Theory to tackle Overlapping Community Detection Problems (OCDPs). The two first algorithms are based on distances to guide the search for partitioning the graph, and these algorithms are called K-fixed and K-adaptive GCF-I. In both algorithms, the fitness functions uses distances to calculate the similarity between the graph nodes, and metrics from the graph theory to perform the split of the graph according to them. However, in the first version of the algorithm (K-fixed GCF-I), the number of communities to partition the graph is fixed (K), and it is given as input parameter of the algorithm. On the other hand, in the afterwards version of the algorithm (K-adaptive GCF-I), a new encoding has been designed to achieve the automatic adaptation of the number of communities (K).

For GCF-I algorithm is necessary that such node has associated a set of features which allow to measure its distance regarding to the rest of graph nodes. Many graphs do not have features associated to their nodes, therefore the third version of the algorithm (GCF-II) tries to find solutions optimizing only metrics based on the own network topology of the graph. For this purpose, a new fitness function has been implemented combining several metrics extracted from Graph Theory. Finally, all these new algorithms have been experimentally evaluated using the Eurovision Contest dataset, that can be considered as a well-known SN.

3.1 Genetic-based Community Finding Algorithm based on distances (GCF-I)

The Genetic-based Community Finding (GCF-I) Algorithm uses a genetic algorithm to detect the best K communities in a graph, where any particular node can belong to different communities or clusters. In this initial approach, a simple version of the algorithm has been developed with a binary encoding using a fixed value for K, which is called *K-fixed GCF-I* or simply K-fixed algorithm. Later, a more complex encoding has been designed to include the value of K in the evolutionary process. This new approach is called *K-adaptive GCF-I*, and it has been compared against the K-fixed version in order to evaluate which of them achieve better overall results. This section describes both versions of the algorithm including the encoding, the genetic operators, and the fitness functions designed.

3.1.1 K-fixed GCF Algorithm (K-fixed GCF-I)

The initial version of the algorithm is based on a standard genetic algorithm, with a binary encoding used to represent the graph partition into communities. The number of the communities to detect (K) is predefined as an algorithm parameter.

3.1.1.1 Encoding

In the first version of the algorithm the genotypes are represented as a set of binary values. Each allele represents a node of the graph, and each chromosome is used to represent a community. Therefore, the chromosome length will be equal to the graph size. This encoding defines a direct relationship between each node in the graph and the allele of the chromosome. In this binary representation the value ”1” means that the node belongs to a community, and the value ”0” the opposite as shown in Figure 3.1.

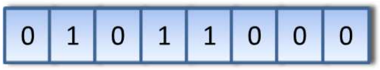


Figure 3.1: A chromosome representing a community. Each allele represents a node of the graph and its belonging, or not, to the current community. In this example, a community built by three nodes from a general graph composed by 8 nodes.

3.1.1.2 Fitness Functions

In this first approach four fitness functions have been implemented, each of them with a different goal. The first one tries to find nodes with a similar rating behaviour (Minimal Distance Fitness), the second one tries to find communities using the clustering coefficient (Maximum Clustering Coefficient Fitness), the third one is similar to the previous one but using the weighted clustering coefficient and, finally, the last fitness function (Hybrid Fitness) combines both strategies (Minimal Distance and Maximum Clustering Coefficient) to find communities with a similar rating behaviour, and whose members are connected between them. These fitness functions can be described as follows:

- **Minimal Distance Fitness (MDF)**. The objective of this fitness function is to find similar nodes to form a community. The evaluation of this fitness function is done using the following criteria:

1. Each node belonging to a community is represented as a vector of attributes. The definition of these attributes depends on the problem to be solved.
2. The average euclidean distance between vectors of attributes within a community is calculated. The fitness calculates distances to be taken into account from peer to peer, between all vectors.
3. The fitness value for the community is the average distance of the values calculated in the previous step. This average distance will be called d_{in} as shown in Figure 3.2.
4. This fitness penalizes those cases where the community has a single node giving it a high value. This is due to the necessity to minimize the MDF to find similar nodes, so the worst values are the highest ones.

- **Maximum Clustering Coefficient Fitness (MCCF)**. The goal of this fitness is to discover communities whose members are strongly connected between them. It is measured through the clustering coefficient, defined as follows: the fitness takes the sub-graph defined by the community and calculates its clustering coefficient. It returns the inverse value, because the genetic algorithm tries to minimize the fitness function.

- **Maximum Weighted Clustering Coefficient Fitness (MWCCF)**. This fitness is similar to the Clustering Coefficient fitness. The main difference between them is the way they consider the community definition. As described in the state of the art, Weighted Clustering Coefficient considers to be stronger those communities whose sub-graphs have higher weight values. It also returns the inverse value, because the genetic algorithm tries to minimize the fitness function.

- **Hybrid Fitness(HF)**. This last fitness function combines both, the Clustering Coefficient and the Distance fitness metrics: it tries to find a set of communities satisfying both conditions previously defined. With this method we try to find strong and similar communities (members that are highly connected between them, and that show similar behaviour). The function defined is a simple weighted function: suppose that $F(x, y)$ is the fitness function, CC the clustering coefficient and d_{in} the distance average between nodes, then the value of HF fitness is calculated as follows:

$$F_i(CC, d_{in}) = w_1 \frac{CC_i}{Max(\{CC_i\}_{i=1}^{K})} + w_2 \left(1 - \frac{d_{in_i}}{Max(\{d_{in_i}\}_{i=1}^{K})}\right) \qquad (3.1)$$

  Where w_i represents the weights given to each fitness: $w_i \in (0, 1)$. The values were fixed experimentally to $w_1 = 0.1$ and $w_2 = 0.9$.

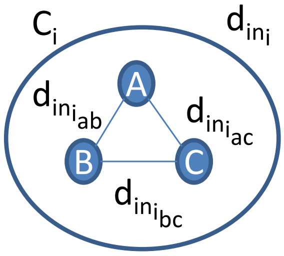
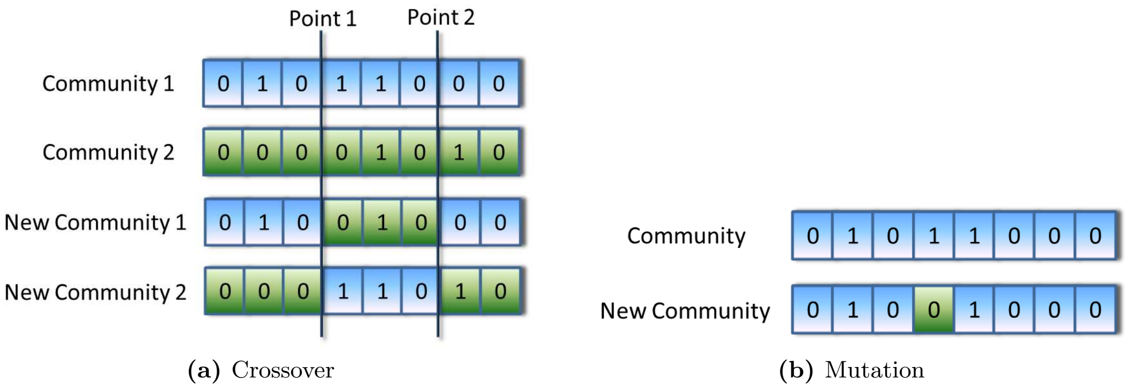


Figure 3.2: Sample illustrating a community and the distance that is calculated by the fitness function of the algorithm. The distance d_{in} represents the average distance calculated between the nodes which belong to a community.

3.1.1.3 The GCF Algorithm

The GCF-I Algorithm with a fixed K works as shown in Algorithm 4. Firstly a random population of communities is generated, containing λ individuals with a length equal to the number of nodes belonged to the graph (line 1). The population evolves using a standard GA until a fixed number of generations, or a convergence value, are reached (lines 2 to 19).

For each generation the individuals are evaluated using the fitness function. When a new offspring is created applying crossover and mutation operators, there is a high probability of loosing the best community (individual). So the algorithm performs an elitism selection method that firstly copies the μ best communities to the new population (line 7 and 8).



(a) Crossover **(b)** Mutation

Figure 3.3: Genetic operators of GCF-I algorithm. The crossover operator (a) used a two point approach, which simply interchanges some elements from Community 1 to Community 2. The (b) example shows how is applied the mutation operator, the fourth allele has been selected, and the bit has been changed using the mutation probability. So this node is now excluded from the community.

To apply the crossover, the algorithm randomly chooses two individual from the μ (lines 10 to 12) performing then a two point crossover operation as shown in Figure 3.3.a. Afterwards, an uniform mutation operator is applied to the two new individuals (lines 13 and 14). This operator is applied to prevent the falling of all solutions into a local optimum of the problem. For a binary encoding, this operator changes the value of the chosen alleles (using probability *mutpb*) between 0 and and 1 or viceversa (see Figure 3.3.b).

Finally, the individuals that are the K-best solutions of the algorithms are selected as output (lines 20 to 37). This entails a *subsumption* process where will prevail the communities that have good fitness and a larger size. Therefore, an individual subsumes another when the subgpraph that represents its community, contains at least all the nodes and connections of the other one (lines 28 to 31). In this process, a ranking of the K best individuals is done by removing those individuals which are contained by another individual of the ranking.

3.1.2 K-adaptive GCF Algorithm (K-adaptive GCF-I)

From the previous version of the algorithm, one of the possible improvements that can be performed is to allow calculating the parameter K (the number of communities which the graph

Algorithm 4: Genetic-based Community Finding Algorithm with a fixed number of communities (K-fixed GCF-I)

Input: A graph $G = (V, E)$ where V is a set of vertices denoted by $\{v_1, \ldots, v_n\}$ and E is a set of edges E denoted by e_{ij} representing whether there is a connection between the vertices v_i and v_j. And positive numbers $ngen$, μ, λ and $mutpb$ are the main GA parameters to be fixed.

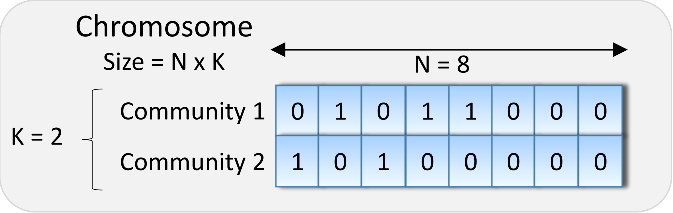
Output: K-best individuals

```
$C \leftarrow InitRamdomPop(\lambda, |V|)$
for $j \leftarrow 1$ to $\lambda$ do
    $F_j \leftarrow Fitness(C_j)$
$i \leftarrow 1$
$convergence \leftarrow 0$
while $i \leq ngen \wedge convergence = 0$ do
    $Cbest \leftarrow SelectNBest(C, F, \mu)$
    $C \leftarrow Cbest$
    for $j \leftarrow \mu$ to $\lambda$ do
        $p1, p2 \leftarrow RandomSel(Cbest)$
        $c1, c2 \leftarrow Crossover(p1, p2)$
        $c1, c2 \leftarrow Mutation(c1, c2, mutpb)$
        $C \leftarrow C \cup \{c1, c2\}$
    $i \leftarrow i + 1$
    for $j \leftarrow 1$ to $\lambda$ do
        $F_j \leftarrow Fitness(C_j)$
    $convergence \leftarrow CheckConvergence(Cbest, C, F)$
$Cbest \leftarrow \emptyset$
$i \leftarrow 1$
$count \leftarrow 0$
$SortedCbest \leftarrow SortNBest(C, F)$
while $count \leq K$ do
    $j \leftarrow 1$
    $stop \leftarrow 0$
    while $j \leq count \wedge stop = 0$ do
        if $SortedCbest_i \subseteq Cbest_j \vee Cbest_j \subseteq SortedCbest_i$ then
            $Cbest \leftarrow Cbest/Cbest_j$
            $Cbest \leftarrow Cbest \cup SelectBigCom(Cbest_j, SortedCbest_i)$
            $stop \leftarrow 1$
        $j \leftarrow j + 1$
    if $j = count$ then
        $Cbest \leftarrow Cbest \cup SortedCbest_i$
        $count \leftarrow count + 1$
    $i \leftarrow i + 1$
return $Cbest$
```

is divided) during the execution of the evolutionary process. For this purpose, the encoding and the fitness function have been modified in the new version of the algorithm.

3.1.2.1 Encoding

In this new approach, the possible solutions can contain groups of communities, and not just an unique community. For this reason, the genotypes (chromosomes) are represented as a set of vectors of binary values. Each allele represents a community that is composed by a set of binary values, one for each node in the graph. This binary vectors are similar to the chromosomes of the previous encoding, the value 1 means that the node belongs to the community, and the value 0 the opposite. The number of binary vectors (communities) contained in the chromosome (group of communities), corresponds to the value of the parameter K, as shown in Figure 3.4.



Community 1	0	1	0	1	1	0	0	0
Community 2	1	0	1	0	0	0	0	0

Figure 3.4: A chromosome representing a group of communities of the graph. Each allele is a particular community where its binary vector represents the nodes of the graph and if they belong, or not, to the current community. So N is the number of nodes contained in the graph. In this example the solution contains 2 vectors representing two different communities, hence the K is equal to 2.

In this new encoding the length of a particular chromosome could vary according to the number of communities (K) in which the graph is partitioned. This variable length of the chromosomes will be adequately managed in the generation process of the initial population. In this process, individuals with different sizes can be created using a initial setup parameter ($maxK$) of the algorithm (from 2 to $maxK$).

In addition, the crossover operator will take into account this parameter to generate correct individuals in the evolutionary process. Therefore, if a chromosome represents a graph partition of K communities, and the number of nodes of the graph is N (length of each allele), the total number of bits contained into this chromosome will be NxK (see Figure 3.4).

3.1.2.2 The Clustering Centroid Fitness Function

The previous encoding only allows to use metrics related to measures of a member belonging to one community, because one individual only represents a single community. However, this new encoding, which can represent a group of communities, makes possible to include measures between the different communities encoding on it.

For this new approach a new fitness function, called **Centroid Fitness (CF)**, have been designed to measure the distance between the community centres belonging to a particular chromosome. This new metric is called d_{out} and it has been represented in Figure 3.5 With this new measure, large distances between centres could be desirable because it represents a bigger gap between clusters or communities.

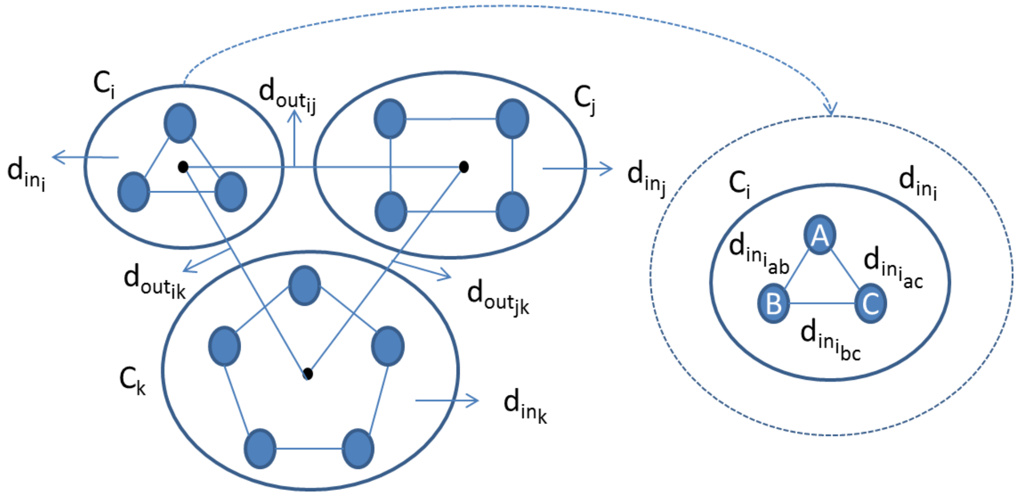


Figure 3.5: Graph sample illustrating three communities and the distances that are calculated using the fitness function of the algorithm. The distance d_{in} represents the average distance calculated between the nodes which belong to a community. The distance d_{out} represents the distance between community centres.

As a result of this new measure, which can be calculated for each individual, a new fitness function which combines the Clustering Coefficient, the distance between nodes (d_{in}) and finally the distance between centres (d_{out}) can be designed. The idea of this new fitness is to find a set of communities that could satisfy all of the previously defined conditions. This new fitness tries to find groups of communities where each community is strongly connected and has similar nodes, but also whose nodes are as different as possible with the rest of communities.

The function defined is a simple weighted function: let $F(x, y)$ be the fitness function, CC the clustering coefficient, d_{in} the distance between nodes, and d_{out} the distance between centres, the value of the new fitness is calculated as follows:

$$F_i(CC, d_{in}, d_{out}) = w_1 \frac{CC_i}{Max(\{CC_i\}_{i=1}^{K})} + w_2 \left(1 - \frac{d_{in_i}}{Max(\{d_{in_i}\}_{i=1}^{K})}\right) + w_3 \frac{d_{out_i}}{Max(\{d_{out_i}\}_{i=1}^{K})} \tag{3.2}$$

Where w_i are the weights given to each fitness: $w_i \in (0, 1)$. The values were experimentally fixed to $w_1 = 0.05$, $w_2 = 0.05$ and $w_3 = 0.9$.

3.1.2.3 The Algorithm

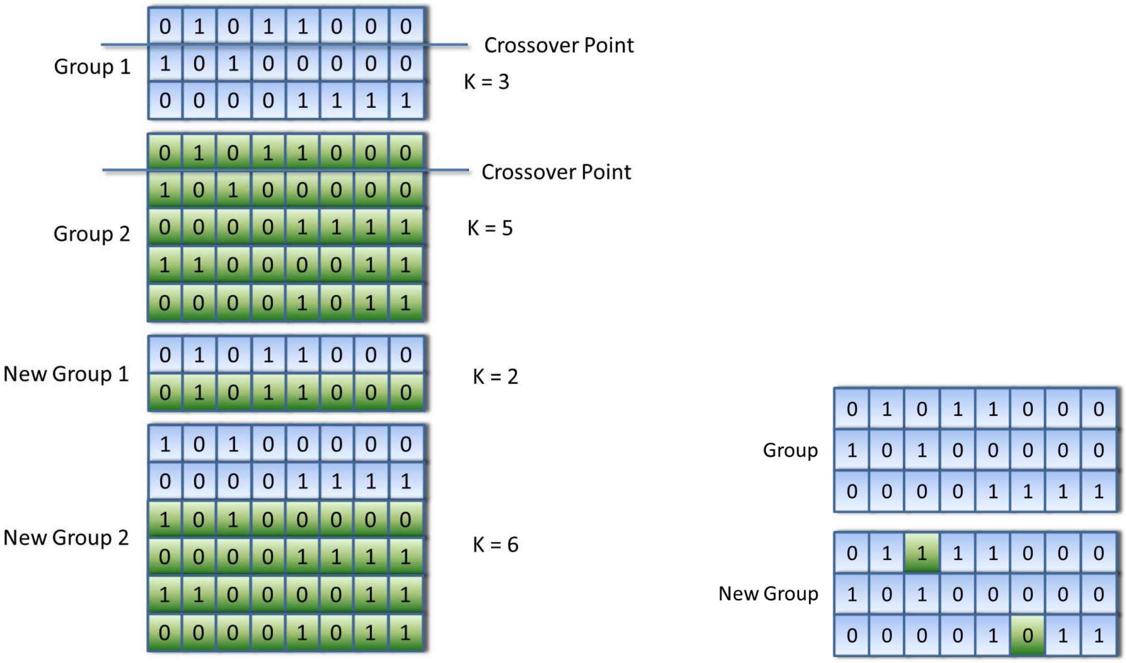
The GCF-I Algorithm with adaptive K evolves using a standard GA. The steps of the process are similar to the previously described in the GCF-I algorithm with fixed K. However, the finally **subsumption** process is no longer required, because any individual represents a group of communities. So in this new approach, the chromosome that has the best fitness function value is selected as a final solution (see Algorithm 5).

Algorithm 5: Genetic-based Community Finding Algorithm with an adaptive number of communities (K-adaptive GCF-I).

```
Input: A graph $G = (V, E)$ where $V$ is a set of vertices denoted by $\{v_1, \ldots, v_n\}$ and $E$ is
       a set of edges $E$ denoted by $e_{ij}$ representing whether there is a connection
       between the vertices $v_i$ and $v_j$. And positive numbers $ngen$, $\mu$, $\lambda$, $mutpb$ and
       $maxK$ are the main GA parameters to be fixed.
Output: Best individual
1  $C \leftarrow InitRamdomPop(\lambda, |V|, maxK)$
2  for $j \leftarrow 1$ to $\lambda$ do
3      $F_j \leftarrow Fitness(C_j)$
4  $i \leftarrow 1$
5  $convergence \leftarrow 0$
6  while $i \leq ngen \wedge convergence = 0$ do
7      $Cbest \leftarrow SelectNBest(C, F, \mu)$
8      $C \leftarrow Cbest$
9      for $j \leftarrow \mu$ to $\lambda$ do
10         $p1 \leftarrow RandomSel(Cbest)$
11         $p2 \leftarrow RandomSel(Cbest)$
12         $c1, c2 \leftarrow CrossoverComm(p1, p2)$
13         $c1 \leftarrow MutationComm(c1, mutpb)$
14         $c2 \leftarrow MutationComm(c2, mutpb)$
15         $C \leftarrow C \cup \{c1, c2\}$
16     $i \leftarrow i + 1$
17     for $j \leftarrow 1$ to $\lambda$ do
18         $F_j \leftarrow Fitness(C_j)$
19     $convergence \leftarrow CheckConvergence(Cbest, C, F)$
20 return Best(Cbest)
```

In the process of random population generation, there is an input parameter ($maxK$) that specifies the maximum number of communities per individual (line 1). For each individual generated, a random value is selected between 2 and the value of $maxK$ parameter, corresponding to the K communities that contains this solution. Therefore, this individual has a size of K binary vectors (with length equals to $|V|$), representing each one a different community of the graph partition. In addition, these binary vectors will be randomly generated too.

Due to the new encoding designed for the algorithm, the genetic operators had to be extended for working with groups of communities as shown in Figure 3.6. To apply the crossover operator, the algorithm chooses a random crossover point. Then, every community preceding this point is copied from both parents to create a new child, and every community succeeding this point is copied to create a second new child (sub-figure a). Once the crossover operator has finished, the mutation is executed. The algorithm randomly chooses some values of the vectors (with a *mutpb* probability) representing the communities, and change their values from 0 to 1 or viceversa (sub-figure b).



Group 1

0 1 0 1 1 0 0 0
1 0 1 0 0 0 0 0
0 0 0 0 1 1 1 1

Group 2

0 1 0 1 1 0 0 0
1 0 1 0 0 0 0 0
0 0 0 0 1 1 1 1
1 1 0 0 0 0 1 1
0 0 0 0 1 0 1 1

New Group 1

0 1 0 1 1 0 0 0
0 1 0 1 1 0 0 0

New Group 2

1 0 1 0 0 0 0 0
0 0 0 0 1 1 1 1
1 0 1 0 0 0 0 0
0 0 0 0 1 1 1 1
1 1 0 0 0 0 1 1
0 0 0 0 1 0 1 1

(a) Crossover

Group

0 1 0 1 1 0 0 0
1 0 1 0 0 0 0 0
0 0 0 0 1 1 1 1

New Group

0 1 1 1 1 0 0 0
1 0 1 0 0 0 0 0
0 0 0 0 1 0 1 1

(b) Mutation

Figure 3.6: Genetic operators for the k-adaptive GCF-I algorithm. In the Crossover example (a), two groups of communities with different K are selected. As shown in the example, the K value is adaptive during the evolution process, where the best graph partitions will survive. In the mutation operation (b), two nodes randomly selected from two different communities have been changed.

3.1.3 Experimental Results

3.1.3.1 Dataset Description

As introduced in the State of the Art chapter, The Eurovision Song Contest has been studied using different Data Mining techniques since the nineties [135, 77, 179]. The main interest in this dataset was the study and analysis of alliances between countries, which had been reflected in form of communities or country clusters. All of these works were able to group the participating countries into blocks of similar behaviour demostrating a community structure in the Eurovision dataset. For this reason this dataset has been selected to carry out the experimental phase of the different versions of GCF algorithm. The data used in these experiments have been extracted from Eurovision's official website [2].

3.1.3.2 Study and comparison of the Eurovision network against a random context

In order to validate that the Eurovision network shows voting patterns, a simple comparison of this graph and a random graph, generated using the same rules applied in the contest, has been carried out. Simulating the contest voting, from each node (participant country) of the graph will be randomly created 10 edges to other 10 different nodes. This simulated graph is called Random Network.

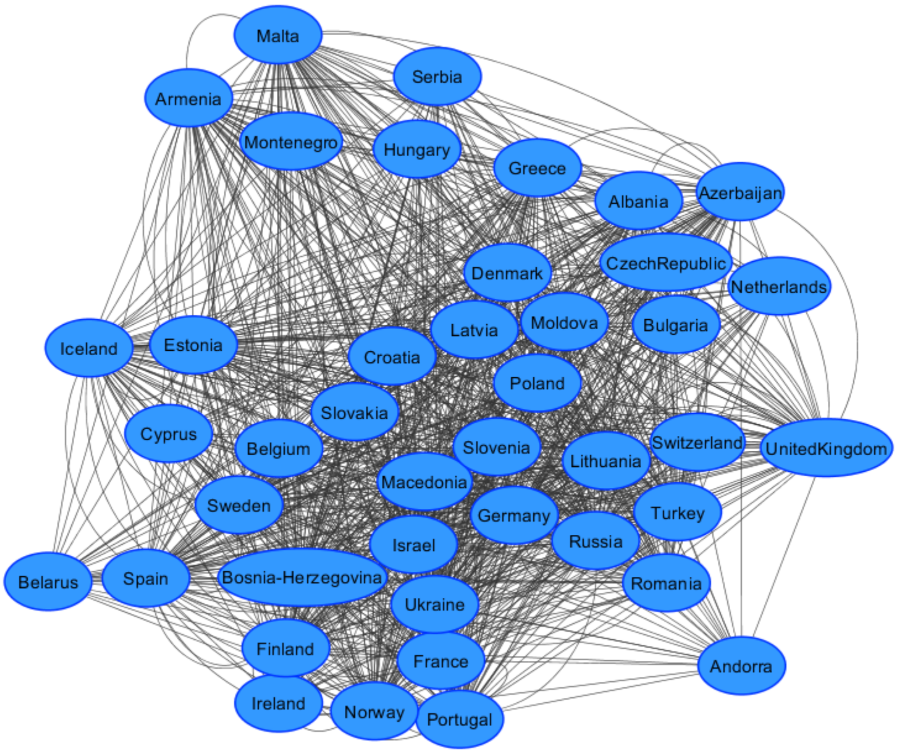


Figure 3.7: Graph representing the votes emitted in the Eurovision Song Contest (2009 year).

The Random Network model assumes that a given country does not favours or penalize other countries and all songs have equal musical quality. So a country X will give points randomly to another ten countries. If, for example, there are N countries then the probability that a country X votes to country Y is given by P = 10/(N-1). Usually, in social networks, two vertices with corresponding edges to a third vertex have a higher probability of being connected to each other. Hence, it may be possible to observe the same effect in the Eurovision network. Therefore, to study this effect it is reasonable to carry out a comparative assessment of the clustering coefficient of both Random and Eurovision networks.

Figure 3.8 shows the clustering coefficients calculated for years between 1992 and 2010. The clustering coefficient for the Eurovision network is always higher than the value achieving to the Random network. It means that the graph distribution of edges is not random, therefore, it can be concluded that exist an “intention of vote” between countries. So, these results provide an evidence that the voting system is not random, and there exist alliances between countries that can be seen as ”communities”.

3.1.3.3 Preliminary analysis of fitness functions

In the previous data analysis using the clustering coefficient, the existence of clusters or communities in the Eurovision graph representation has been confirmed. Specifically, the 2009 year has the greatest difference on the clustering coefficient, meaning that this year contains communities with stronger connectivity. Hence, this year have been chosen to perform an initial study for all the fitness functions designed and described in previous section.

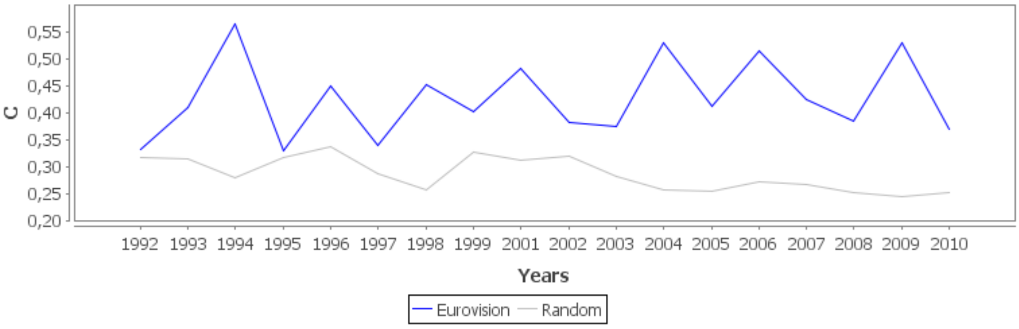


Figure 3.8: Comparative assessment of the CC between the Eurovision network, and the random graph simulated for each year.

Table 3.1 shows the set-up for the K-fixed and K-adaptive versions of the algorithm used in the experiments. All the parameters have been experimentally obtained.

Algorithm	K-fixed	K-adaptive
Mutation probability	0.2	0.03
Generations	2500	500
Population size (λ)	3000	1000
Selection criteria (μ)	300	100
K value	6	-

Table 3.1: Parameters values for both versions of GCF-I algorithm.

To compare the results obtained by the different fitness functions, the following measures (which have been previously defined) are considered:

- d_{in}: It provides information related to the node similarity within communities.
- CC: It provides information related to the inner connections of the communities.
- d_{out}: It provides information related to the distances between the centroids of the communities.

Then, to analyse in detail the results achieved, this preliminary study has been divided in two parts, one for each version of the algorithm (K-fixed and K-adaptive) as shown below:

1. *Fitness Function Analysis for the K-fixed algorithm.*

   Table 3.2 shows the values of the clustering coefficient and the distances (d_{in} and d_{out}) obtained for each fitness function using the K-fixed version of the algorithm. As shown in this table, in terms of distance measures, the hybrid fitness (HF and WHF) functions greatly improve the results. The distance between centres (d_{out}) increases dramatically from the MCCF and MWCCF functions to the hybrid ones. Therefore, the communities found are far from each other, and they can be better differentiated. The intra-cluster distance (d_{in}) obtains lower values, meaning that the found communities have more similar members. Finally, the clustering coefficient takes similar and very high values in all of the analysed cases.

	MCCF	**MWCCF**	**HF**	**WHF**
d_{in}	21,15	21,56	**18,2**	19,02
d_{out}	5,4	6,39	11,26	**11,99**
CC	**1**	**1**	0,9	**1**

Table 3.2: Values for clustering coefficient and distances d_{in} and d_{out} obtained using weighted and unweighted fitness functions for K-fixed algorithm (2009 dataset).

On the other hand, the addition of weights to the functions improves the distance (d_{out}), and the clustering coefficient fitness functions. However, it worsens the distance (d_{out}) in the hybrid fitness. Based on these experimental results it can be concluded that overall the hybrid approaches perform better, and the weights in the clustering coefficient do not greatly affect the outcome.

2. *Comparison of fitness functions for K-fixed and K-adaptive algorithms.*

   Using the previous experimental conclusions, the next experiments have been executed using the K-adaptive algorithm, and the clustering coefficient (without weights). Figure 3.9 shows the experimental results comparing both versions of the algorithm. Fitness functions labelled with an asterisk represent the results for the K-adaptive algorithm.

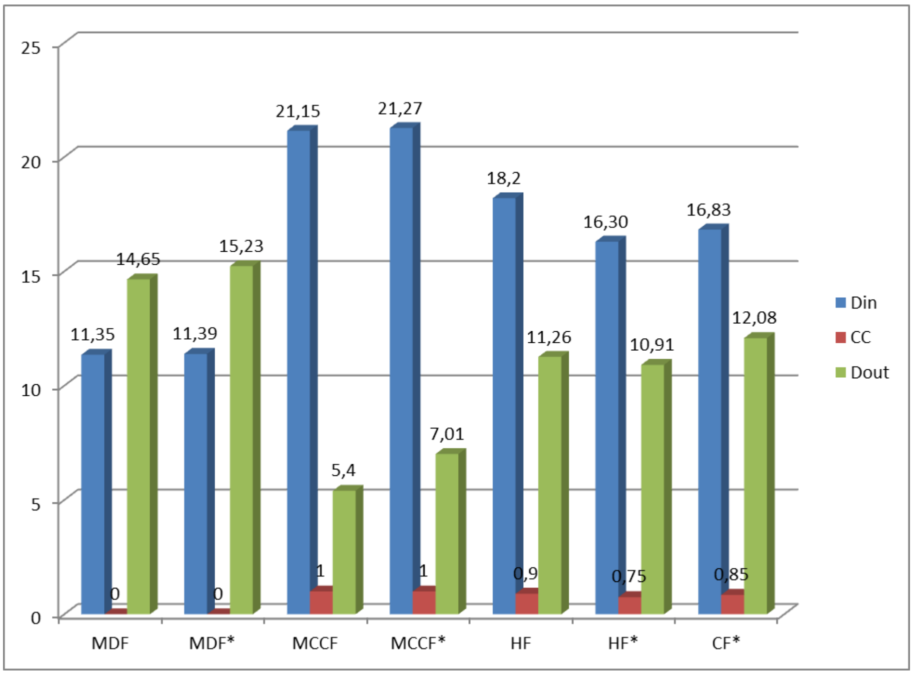


Figure 3.9: Values for the clustering coefficient and the distances d_{in} and d_{out}, obtained using the designed fitness functions with both versions of the algorithm. The fitness functions labelled with an asterisk show the values for the K-adaptive algorithm.

As shown in this figure, the first two fitness functions, (MDF and MDF^{*}), obtain the minimum (d_{in}) distance and the maximal d_{out} distance, but the value of the CC is 0 in both cases. It means that the members of the communities are not connected between them. In the next two functions ($MCCF$, MCCF*) the opposite situation is noticed. The maximum possible value of CC is reached, but the distance measures are dramatically worse.

Both approaches have been combined into new hybrid fitness functions (HF, HF^*) that try to find new communities with better values for all the considered measures. Figure 3.9 shows that the distance between centres (d_{out}) and the intra-cluster distance (d_{in}), take values lying between the first and second functions. Finally, the clustering coefficients (0,9 and 0,75 respectively) are closer to the values obtained by the second fitness functions, that obtain the maximum possible value (1).

The last fitness function considered, the centroid fitness function (CF), obtains similar results for CC and d_{in} values and improves the d_{out} distance. This expected result comes from the own definition of this function, that uses the distance between centroids to determine how to build the community.

Finally, all the experimental results from these fitnesses are compared for both versions of the algorithm. It can be noticed that the K-adaptive algorithm obtains similar or better results than the K-fixed algorithm in all the studied cases. Therefore, the CF function has been selected to experimentally test the new community finding approach against other classic community finding algorithms.

3.1.3.4 Comparative assessment of algorithms

In this section, two classic and well-known algorithms for community detection, CPM and EBC, are compared against the two versions (K-fixed and K-adaptive) of the GCF-I algorithm. In order to carry out this comparative assessment, 10 different graphs have been chosen, representing the contest voting from different years. Taking into account the history of Eurovision, the periods used in the comparison must be carefully picked to prevent noise produced by the many evolutions of the voting system. Therefore, the periods which have been selected are:

- 1992-1996: Jury-based voting system was used exclusively.
- 2004-2008: Televoting was used exclusively, as well as having a semi-finals round.

Because the period from 1997 to 2003 saw a slow adoption of the *televote* system, the data is not representative, and it has not been included in this study.

Figure 3.10 shows that the d_{in} measure is minimized by both genetic algorithms, however the first version of the algorithm obtains better results. The new approach has closer results and there is a big gap between the genetic algorithms and EBC or CPM. It means that the nodes of the GCF-I algorithms have more similar nodes than the EBC and CPM algorithms.

Figure 3.11 shows the CC measure results, and as it can be observed, its value is maximized by both genetic algorithms. In this case the new genetic algorithm approach, using the K-adaptive version, obtains the best results, followed by the K-fixed version of GCF. The EBC

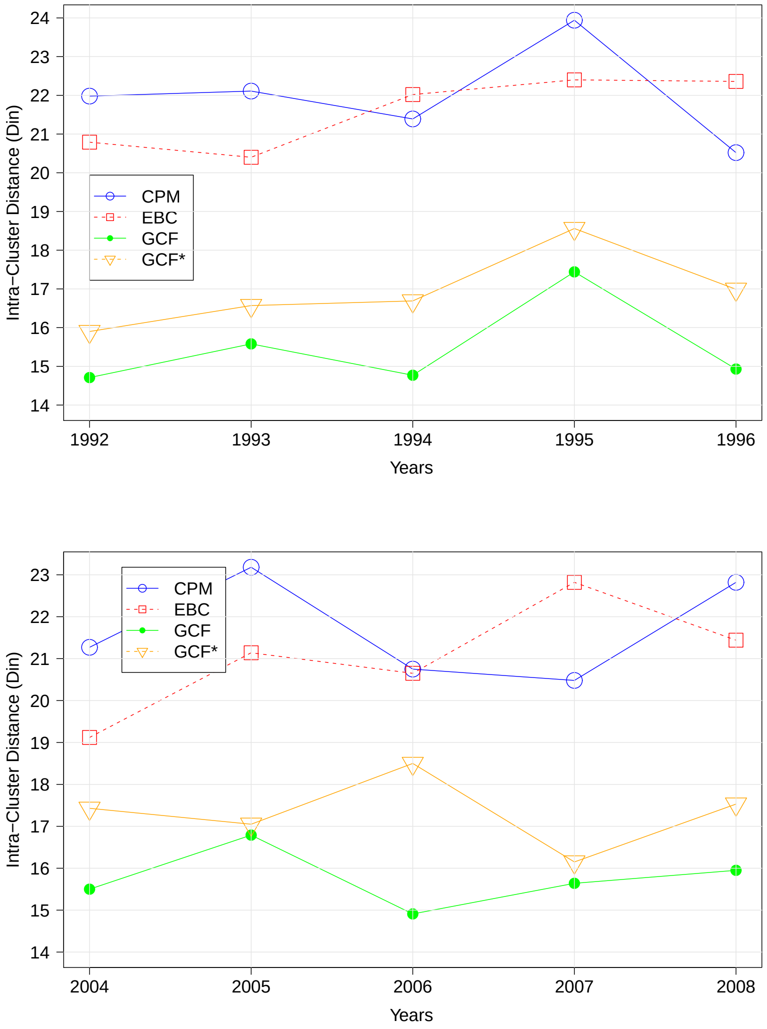


Figure 3.10: d_{in} comparison of a dataset collection extracted from the Eurovision Song Contest.

and CPM algorithms obtain the worst CC results, meaning that there are fewer connections between nodes within communities.

Regarding the d_{out} measure, Figure 3.12 shows that it is maximized by both genetic algorithms. As in the previous case, the K-adaptive approach obtains the best results. It was one of the original goals of the algorithm improvements. The difference observed in cluster centroid distance, between the results obtained by K-fixed version and those through by EBC and CPM algorithms, is not too significant. Nonetheless, the K-adaptive version always improves that value.

Finally, from an overall analysis of all the studied metrics, it can be noticed that the new K-adaptive GCF algorithm improves the results in all the cases comparing to the results obtained through the classic methods for community detection.

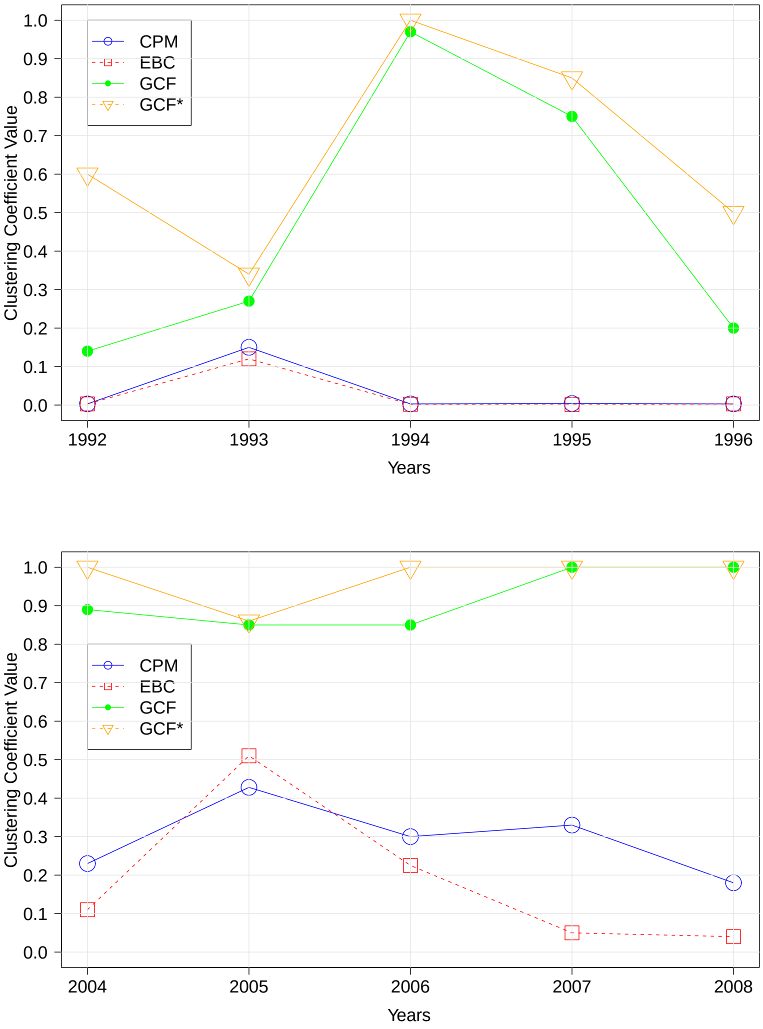


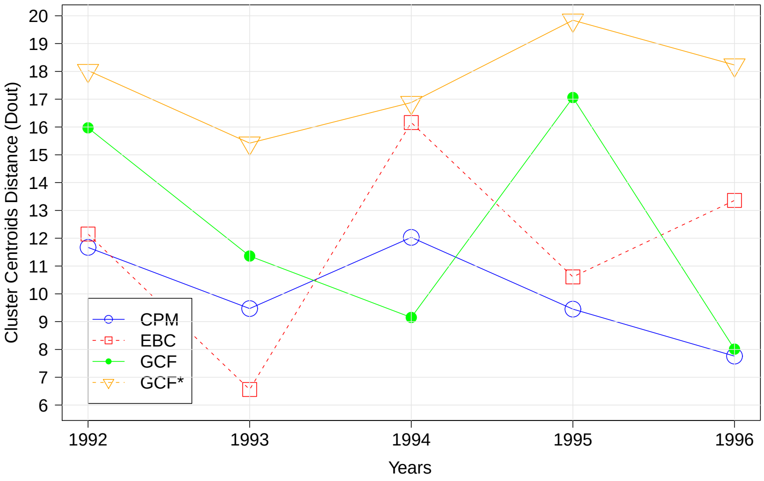
Figure 3.11: *CC* comparison of a dataset collection extracted from the Eurovision Song Contest.

3.1.3.5 Community Interpretation

In this section a human interpretation is given to the results obtained by the different algorithms. For this purpose, one particular year is selected (2006) to make a more detailed analysis. The found communities, for each algorithm, have been plotted in a geographical context to study the establishment of alliance between countries.

The first map plots the results applying the CPM algorithm (Figure 3.13). As shown in this figure, there are big communities with great overlapping, where overlapped countries are marked in bold. Analysing the neighbourhood between countries, a high correlation between neighbouring countries, and their membership to the same communities, can be appreciated. This effect can be appreciated in some subsets of the detected communities as shown below:

- *Baltic States* (Lithuania, Latvia, Estonia) belonging to the community 3.



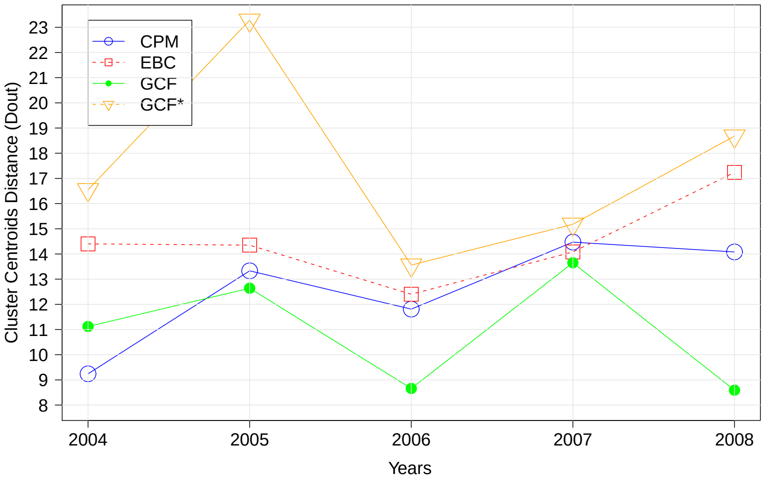


Figure 3.12: d_{out} comparison of a dataset collection extracted from the Eurovision Song Contest.

- *Nordic Countries* (Norway, Sweden and Finland) belonging to the community 4.

- *Balkan Countries* (Macedonia, Albania, Serbia, Bosnia and Herzegovina, Croatia and Slovenia) belonging to the community 4.

- *Countries from the former Soviet Union* (Russia, Belarus, Ukraine and Armenia) belonging to the community 7.

These results show that several of the communities identified by CPM algorithm comprise neighbouring countries sharing common cultural roots, and even the language in some cases. However, regarding the metrics used to evaluate the algorithms in the previous section, this algorithm obtains low values related to d_{in} and CC. It means that the country voting pattern is not too similar, and in addition there is a sparse connectivity between the member countries of the communities.

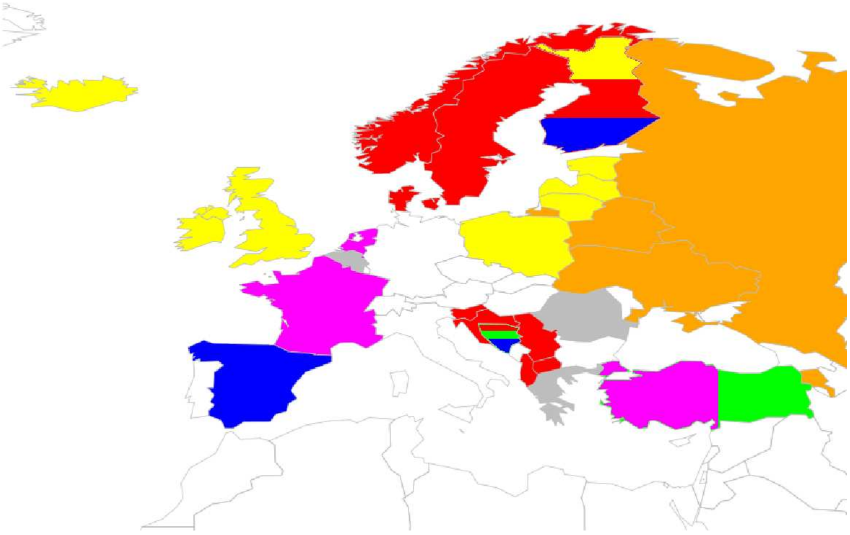


Figure 3.13: CPM Cluster Results for Eurovision 2006. The communities are: 1[Spain, **Bosnia and Herzegovina** and **Finland**], 2[France, Netherlands and **Turkey**], 3[Iceland, Ireland, United Kingdom, Poland, Lithuania, Latvia, Estonia and **Finland**], 4[Norway, Sweden, **Finland**, Macedonia, Albania, Serbia, **Bosnia and Herzegovina**, Croatia and Slovenia and Denmark], 5[Belgium, Romania and Greece], 6[**Turkey**, **Bosnia and Herzegovina**] and 7[Russia, Belarus, Ukraine, Armenia].

On the other hand, the communities resulting from the K-adaptive GCF-I algorithm are smaller as shown in Figure 3.3. This is expected if we consider that the new algorithm tries to find communities whose members are highly connected between them, and also have similar characteristics. Regarding to the neighbourhood of the countries, there is still a high correlation between neighbouring countries and their membership to the communities. Although in this case, the subset of neighbouring countries are smaller than in CPM results (for example communities 2 and 4). However, related to the evaluation metrics, as mentioned previously, all the metrics have been improved by the new approach. Therefore, the K-adaptive GCF-I algorithm is able to achieve better outcomes in terms of quality, minimizing the distance between the elements which belong to a community, and maximizing the cluster centroid distance.

Finally, analysing in detail the found members of the communities identified by both CPM and K-adaptive GCF-I algorithms, an important issue appears immediately. Each community generated by the genetic algorithm is contained, or partially contained, in a community generated by the CPM algorithm. Table 3.3 shows for each community obtained, using the new genetic algorithm, the related CPM community which contains it. For example, that is the case for the community formed by Sweden, Norway and Finland or the formed by Lithuania, Latvia and Ireland shown in this table. Both communities are fully contained in the related communities detected by CPM algorithm.

Taking into account all the experimental findings, it can be concluded that this new approach is able to reach better results than the other classical approaches studied. The K-adaptive GCF-I algorithm finds communities that have an appropriate size, reduced overlapping, and closer distances between the nodes belonging to the communities detected.

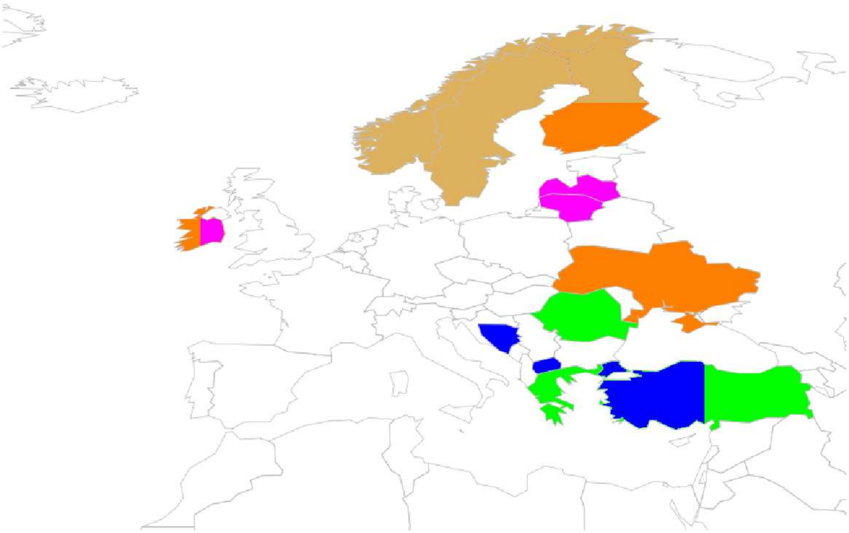


Figure 3.14: K-adaptive GCF-I Cluster Results for Eurovision 2006. The communities are: 1[**Ireland**, **Finland**, Ukraine], 2[**Turkey**, Macedonia, Bosnia and Herzegovina], 3[Lithuania, Latvia and **Ireland**], 4[Sweden, Norway and **Finland**], 5[Greece, **Turkey** and Romania].

K-adaptive GCF-I	CPM	Overlap (%)
Ireland, **Finland**, Ukraine	Iceland, **Ireland**, United Kingdom, Poland, Lithuania, Latvia, Estonia, **Finland**	66%
Turkey, **Macedonia**, **Bosnia and Herzegovina**	Norway, Sweden, Finland, **Macedonia**, Albania, Serbia, **Bosnia and Herzegovina**, Croatia, Slovenia, Denmark	66%
Lithuania, **Latvia**, **Ireland**	Iceland, **Ireland**, United Kingdom, Poland,**Lithuania**, **Latvia**, Estonia, Finland	100%
Sweden, **Norway**, **Finland**	**Norway**, **Sweden**, **Finland**, Macedonia, Albania, Serbia, Bosnia and Herzegovina, Croatia, Slovenia, Denmark	100%
Greece, Turkey, **Romania**	Belgium, **Romania**, **Greece**	66%

Table 3.3: Comparative analysis between the communities detected by K-adaptive GCF-I, and the resulting communities detected by CPM for Eurovison 2006 dataset. The countries of the communities detected by K-adaptive GCF-I that are contained in a CPM community are marked in bold.

3.2 Community Finding Algorithm based on network topology metrics (GCF-II)

In the previous version of the algorithm, the fitness functions used distances (intra-cluster and inter-cluster distances) to guide the search for partitioning the graph. To calculate distances between the graph nodes, it is necessary that such nodes have associated a set of features which allow to measure them. Many graphs do not have features associated to their nodes, therefore it would be interesting that the algorithm tries to find solutions optimizing metrics based on the own network topology of the graph. For this purpose, a new fitness function has been implemented combining several metrics from the network topology in order to tune up the community sets detection.

3.2.1 The Fitness Function based on network topology metrics

The new fitness function designed combines metrics from the Network Topology area. It tries to find a set of communities where their members are highly connected between them, and they have similar behaviour. To combine several network metrics in a single fitness function, a weighted function based on these metrics chosen has been designed. The measures used in the fitness function, and their weights, can be changed in the algorithm setting. This function is calculated as follows:

$$F = \sum_{i=1}^{n} w_i * Metric_i \tag{3.3}$$

Where $Metric_i$ is a network metric extracted from the graph theory to guide the evolutionary algorithm, and w_i are the weights given to each metric: $w_i \in (0, 1)$. In order to compute the metrics chosen and their weights, a preliminary analysis of them is done in the next section. This fitness function will be based on the following metrics: clustering coefficient, density, centralization and heterogeneity. Previous metrics derive from the graph theory, and all of them were introduced in the chapter related to the state of the art.

3.2.2 Analysis of network metrics

The previous data analysis of Eurovision graphs performed using the CC confirmed the existence of clusters or communities in these dataset. Specifically, the graph representing the 2009 year had the greatest difference in the CC values. It means that this year contains communities with a stronger structure than the rest of the analysed years. Hence, this year has been selected to perform a detailed study for different network measures which can be later used to tune up the new fitness function designed.

Table 3.4 shows the parameter setting of the genetic algorithm used throughout the analysis of the network metrics used in the fitness function. These parameters were obtained experimentally by performing several tests with different range of values.

Parameter	**Value**
Mutation probability	0.03
Generations	100
Population size	500
Selection criteria $(\mu + \lambda)$	$500 + 50$

Table 3.4: Parameter setting of the Genetic Algorithm for community detection based on networks metrics.

The concept of good partition for a set of elements is sometimes quite subjective. There are two objective functions in clustering literature to measure the cluster quality: intra-cluster distance (elements within a cluster should be close) and inter-cluster distance (elements from different clusters should be away). These two distance measures have been calculated to compare the results obtained by each network measure used in the fitness function. The values obtained for each measure are shown in Table 3.5.

Metric	Com.	Countries	Intra Dis.	Inter Dis.
Density	1	Denmark Greece	**19,25**	**15,79**
	2	Norway Romania		
	3	Albania Russia		
Centralization	1	Belgium Ireland Ukraine Hungary	21,32	14,66
	2	Albania Serbia		
Heterogeneity	1	Norway Sweden Armenia Albania Moldova Israel Denmark Finland Ukraine Turkey Azerbaijan	20,72	4,92
	2	Norway Croatia Estonia Sweden Albania Moldova Israel Denmark Finland Lithuania Ukraine Turkey Germany Azerbaijan		
	3	Croatia Sweden BosniaandHerzegovina Armenia Albania Malta Russia Finland Romania Lithuania Ukraine Germany Azerbaijan		
	4	Croatia Sweden BosniaandHerzegovina Malta Moldova Denmark Finland Lithuania Ukraine Iceland Germany		
	5	Norway Armenia Moldova Israel Denmark Finland Spain Iceland Turkey Azerbaijan		
	6	Estonia BosniaandHerzegovina Albania Moldova Israel Denmark Ukraine Iceland Germany Azerbaijan		
CC	1	Norway Ukraine Azerbaijan	21,27	7,01
	2	Moldova Ukraine Azerbaijan		
	3	Russia Ukraine Azerbaijan		

Table 3.5: Comparative assessment of network measures for the graph representing the contest votes of the 2009 year.

The algorithm goal is to discover communities consistent internally, but clearly different from the rest. Members within a community should be as similar as possible (lower intra cluster distance), and members in one community should be as dissimilar as possible from members in other communities (higher inter cluster distance). As shown in Table 3.5, the fitness function based on **Density** metric obtains the better results (with a lower intra cluster value and the higher inter cluster value). However, these detected communities are small and no-overlapping. On the other hand, using the **Heterogeneity** metric as fitness function, the best results have been achieved in terms of the size and the overlapping of the communities. In this case the communities detected are bigger with great overlapping. Therefore, both measures have been combined in the new weighted fitness function trying to detect communities with better results considering all the features.

Once the best network measures for the fitness function have been selected (Density and Heterogeneity), it is necessary to perform the estimation related to the best weight for each measure within the function. For this purpose, a comparative assessment of weights for both measures is carried out as shown in Figure 3.15. Analysing the results, with Density equals to 0.9 and Heterogeneity equals to 0.1 (0.9D-0.1H) the intra cluster distance is minimized and the inter cluster measure takes the highest values. The intra cluster distance progressively is worse while its value increases. Therefore, these values have been finally selected to fix the fitness function.

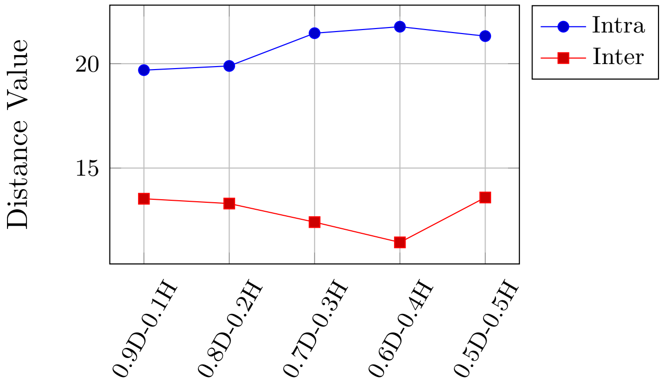


Figure 3.15: Comparative evaluation of weights for the network metrics used as fitness function. The x-values represent the weight considered for Density (i.e 0.9D) and for Heterogeneity (i.e 0.1H).

3.3 Comparative assessment of algorithms

Finally, the results obtained for the two different approaches of GCF algorithm, one based on distances (GCF-I), and the other based on network measures (GCF-II), have been compared against the results of CPM algorithm. For this purpose the graphs chosen for these experiments are the same as in the experimental phase of GCF-I algorithm, representing the contest voting before and after the establishment of the televoting system. The analysis of the experimental results for GCF-I shown, that the K-adaptive version is be able to detect better communities. So, this version of the GCF-I algorithms is used for the following experiments.

Modularity is one of the most used, and best known, quality measures to evaluate graph clustering techniques, so this metric will be used to carry out the comparative assessment. The standard definition of this metric is only used to assess disjoint partitions. Because all of the algorithms applied in this comparative assessment detect overlapping communities, the definition of the Modularity extended by Nicosia et al. [134] for overlapping partition is used. Using the standard definition, a higher positive value corresponds to a better partitioning, being 1 the greatest possible value.

As shown in Figure 3.16 the new version of the algorithm (GCF-II) improves the results in all of the years. Before the establishment of the televoting system the improvement of this evolutionary approach is greater than the rest of the algorithms. In these years the number the of nodes and edges is lower, due to the fact that the semi-final rounds began in 2004, and therefore, more participant countries can vote in the final round too. The results show how this change in the network topology affects to the quality of solutions detected by the three algorithms, although GCF-II still reach the best results. However, CPM and GCF-I present very similar values on Modularity before the Televotin period, during this period, they begin to take differentiated values. Therefore, the main conclusion that can be obtained is that these algorithms are more dependent on the network topology.

Analysing the results of both GCF versions (I and II), it can be concluded that the best results have been achieved finding communities only based in network metrics, without the necessity of measuring distances related to the node features. Therefore the new algorithm version (GCF-II) can be used to discover communities for a greater variety of types of graphs.

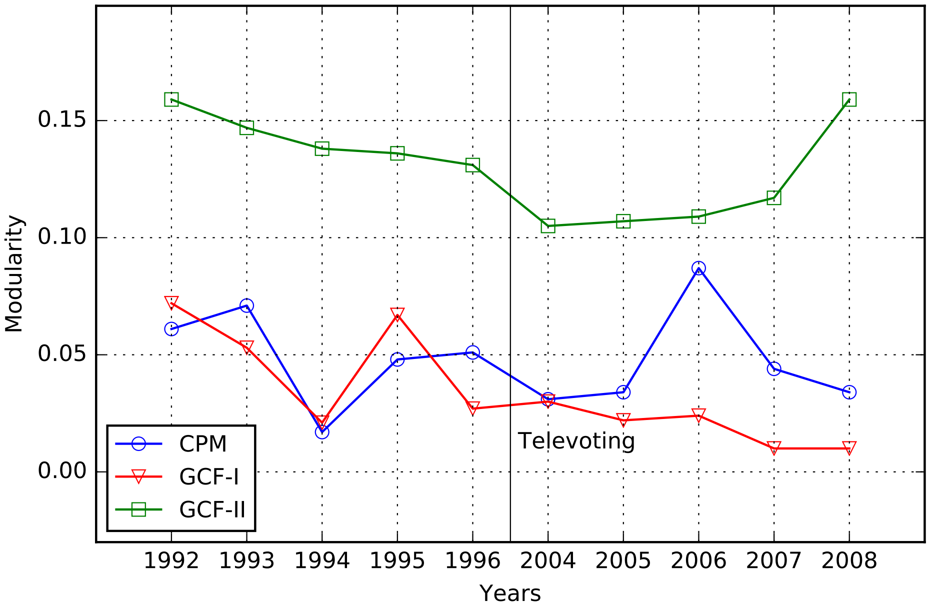


Figure 3.16: Comparative assessment of algorithms using Modularity as evaluation metric.

Additionally, the communities detected for a specific year are plotted in a graph representation. This representation provides a better appreciation of the community structure and size to analyse them. To carry out this analysis, the 2006 year has been selected, because is the central year of the televoting period where this new system has already been totally established. Analysing these features there are several remarkable aspects. As shown in Figure 3.17, CPM algorithm splits the graph into a higher number of communities, which are big and with slightly overlapping. However, as seen above this algorithm obtains the worst values related to modularity, meaning that the finding communities are less structured than those detected by the evolutionary approaches (K-adaptive GCF-I and GCF-II).

On the other hand, the resulting communities for the genetic algorithms are smaller and present more overlaps between them, as shown in Figure 3.18 and Figure 3.19. In particular, the K-adapted GCF-I algorithm discovers 5 communities where all of them contain a node belonging to two different communities. The size of all the communities is 3 which is a size quite small. The algorithm tends to find communities according to the triangles contained in the graph. The fitness function of this genetic algorithm uses the distances between centroids and the distance between the nodes belonging to a group to guide the community detection. This result could be expected, because the encoding designed does not considerer restrictions on the community size. Therefore, it is more likely to find small communities containing very similar nodes, and whose community centres are separated. The distances between the community centres decrease according to their increased size. Finally, regarding the modularity metric, K-adapted GCF-I algorithm improves the results achieved by CPM method, finding a graph partition more structured from the point of view of the network connectivity.

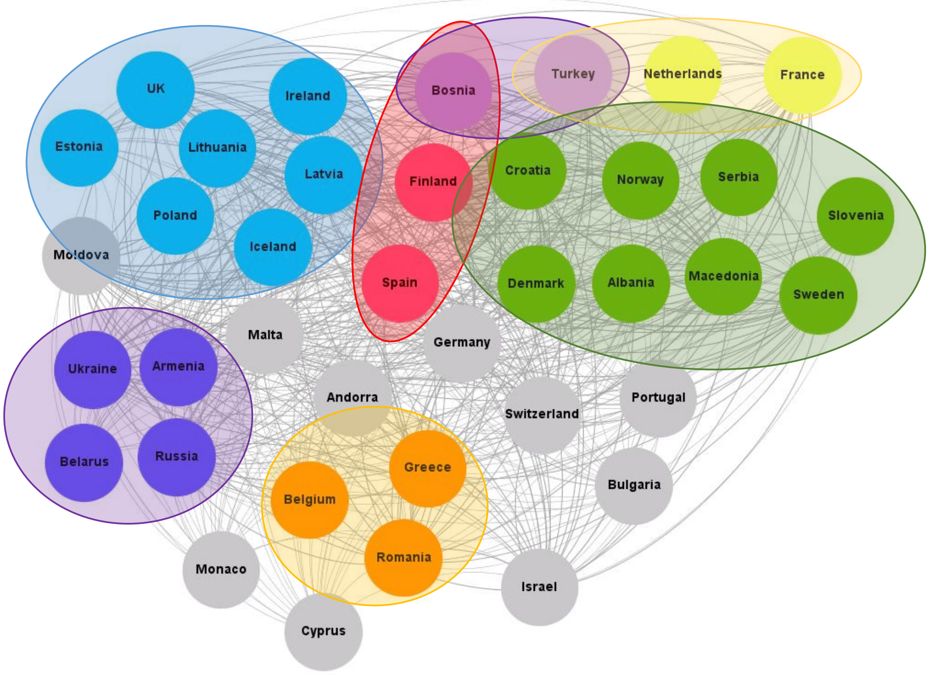


Figure 3.17: CPM communities detected for 2006 year. The graph has been partitioned into 7 communities containing 2 overlapping nodes. Nodes not assigned to any community are in grey.

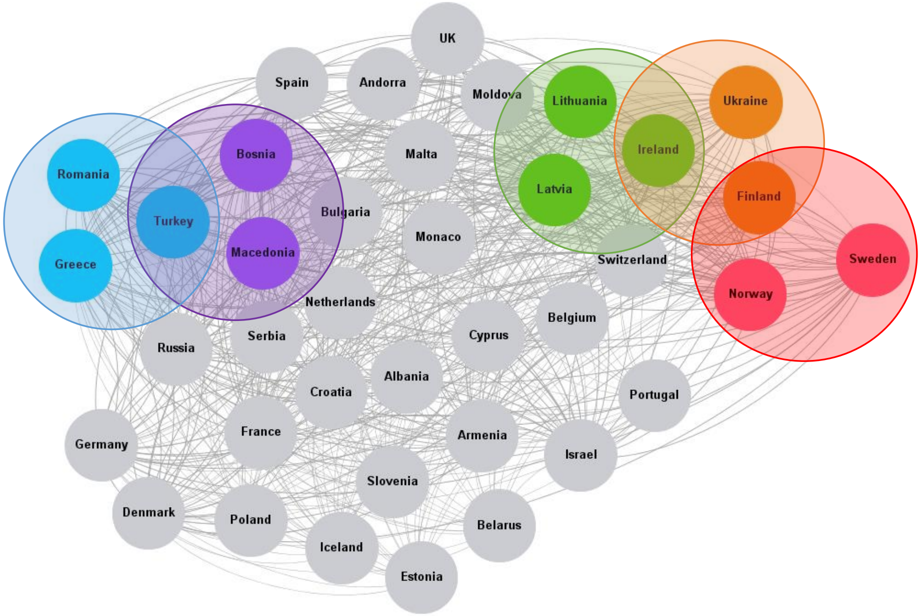


Figure 3.18: GCF-I communities detected for 2006 year of Eurovision Song Contest. The graph has been partitioned into 5 communities containing 3 overlapping nodes. Nodes not assigned to any community appear in grey.

Comparing both evolutionary approaches, the new algorithm (GCF-II) obtains the same modularity value (see Figure 3.16), so the structure level of the finding communities is similar.

However, there are more communities detected using this new approach and with have larger size, as shown in Figure 3.19. These results show that the new approach is able to reach better overall outcomes. This new fitness function of this approach has been inspired by the network topology analysis, and it is only based on the use of network measures to guide the search. It means that the GCF-II algorithm is be able to find good solution to more types of graphs, and not only those that have features associated to their nodes. These results demonstrate that the new algorithm performs a better partition of the graph regardless the votes associated to each node.

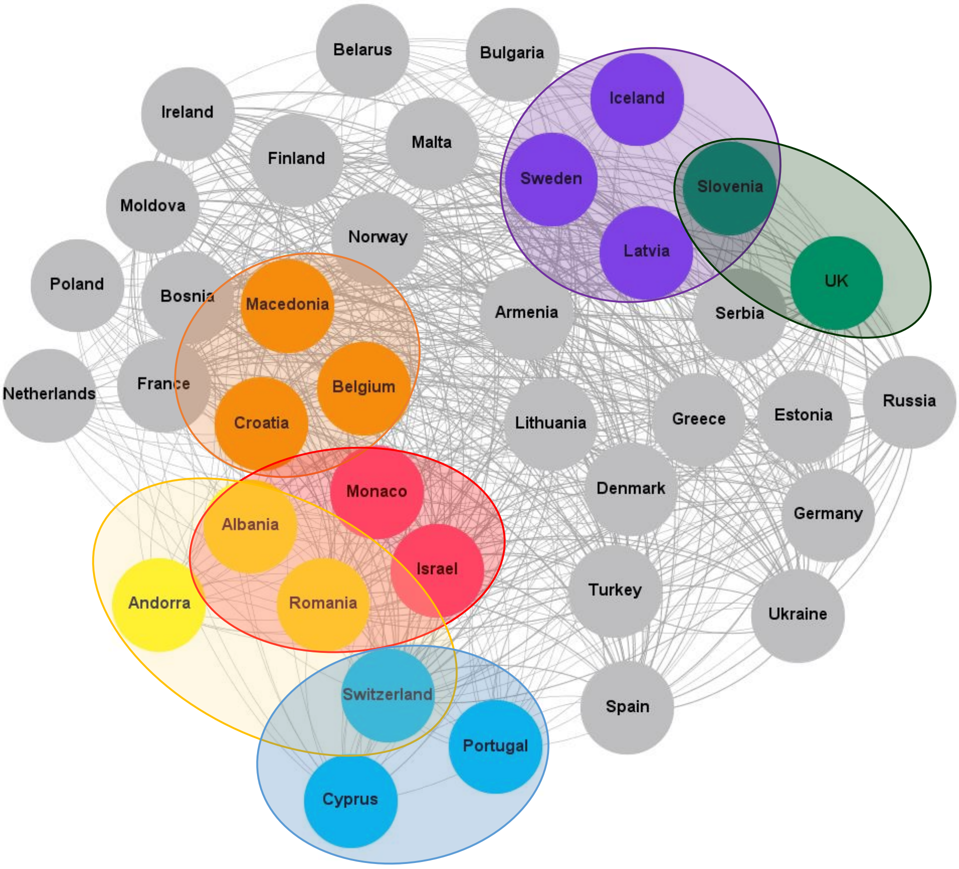


Figure 3.19: GCF-II communities detected for 2006 year of Eurovision Song Contest. The graph has been partitioned into 6 communities containing 4 overlapping nodes. Nodes not assigned to any community are in grey.

Chapter 4

Multi-Objective Genetic Approaches for Overlapping Community Detection

"Science and everyday life cannot and should not be separated.
Science, for me, give a partital explanation for live.
In so far as it goes, it is based on fact, experience and experiment."

- Rosalind Elsie Franklin

This chapter presents two new Multi-Objective Genetic Algorithms for Overlapping Communities Detection (MOGA-OCD), which have been designed to use measures based on the Network Connectivity as optimization criteria of the fitness function. The main difference between these algorithms is based on the encoding used to represent the individuals. The encoding for the first algorithm is node-based, where the alleles represent the nodes of the graph. Whereas, the second algorithm uses an edge-based encoding, where each allele is related to specific edge of the graph. In both algorithms, the value of K (number of communities to find in the graph) has been directly encoding as part of the chromosome, so the evolutionary process will generate the final number of communities as part of the solution.

On the other hand, the fitness function implemented in both algorithms, optimizes two different objective functions; the first one that is used to maximize the internal connectivity of the communities, and the second one that is used to minimize the external connections to the rest of the graph. To select the most appropriate metrics for these objective functions, a comparative assessment of several connectivity metrics has been carried out using a real-world network. Finally, the two new algorithms have been compared among themselves, and evaluated against other well-known algorithms from the state of the art in CDP. The experimental results show that these new algorithms overall improve the accuracy and quality of current approaches in CDP, showing its effectiveness as a new method for detecting structured communities.

4.1 Node-based MOGA-OCD Algorithm

This section introduces the new Multi-Objective Genetic Algorithm that has been designed and implemented in this thesis to handle the OCD Problem. Following the same encoding type used in the algorithms presented in the previous chapter, a node-based representation of the individuals has been designed for this algorithm. The next subsections describe in detail, this node-based encoding, as well as the fitness function, and how the algorithm works.

4.1.1 Node-based Encoding

In the new algorithm, the genotypes (chromosomes) are represented as vectors of integer values. Each chromosome is used to encode the partition of the graph into group of nodes, where overlaps between the groups may exist. Their length is variable, and each allele represents a node belonging to a community. Therefore, the value of these alleles is an integer value between 0 and $n - 1$, being n the number of nodes of the graph. Finally, there are some alleles with a special value, equals to -1, which are used to identify the boundaries of the communities as shown in Figure 4.1.

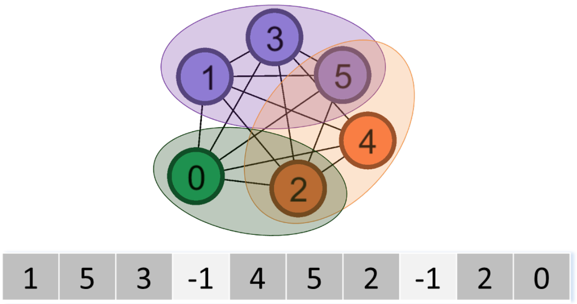


Figure 4.1: Chromosome representing a group of communities in a graph with 6 nodes. The alleles with -1 value mark the boundaries of the communities. The rest of the alleles represents a particular node in the graph. In this example the solution contains 3 different communities with two overlapping nodes (2 and 5).

4.1.2 Fitness Function

The individual evaluation is one of the most important step in any evolutionary algorithm, this evaluation is computed in terms of a fitness function measuring the quality of the found solutions. As it has been previously described, a *'good'* community should be cohesive, compact, and internally well connected. Moreover, regarding to the rest of the graph, this community should be well separated having relatively few external connections.

Taking into account both features, the fitness function implemented for the new algorithm optimizes **two objective functions**, to identify internally connected groups of nodes and with sparse externally connections to the rest of nodes. For this purpose, different topology metrics of a graph, which have been previously described in State of the Art chapter, are used as an optimization criteria for these objective functions:

1. The first objective function uses **internal measures** extracted from the graph theory (Density, Triangle Participation Ratio, Clique Number and Clustering Coefficient), maximizing the internal connectivity of the communities detected by the algorithm.

2. The second objective function is based on **external measures** (Expansion, Separability and Cut Ratio) to identify separate communities in relation to each other. So, when the Separability metric is used, this objective will be maximized too. However, in the case of Expansion and Cut Ratio metrics, to find separate communities, these objectives must be minimized. The specific metric used by each objective function can be changed in the algorithm settings as an argument to execute it.

4.1.3 Algorithm Description

In the new community detection algorithm, the population evolves according to a Multi-Objective Genetic Algorithm, where the found solutions (individuals) are evaluated using a Pareto-Optimality Frontier (POF) of both optimization criteria described in previous section.

As shown in Algorithm 6, the first generation creates randomly a population of λ individuals representing the partition of the graph into overlapping communities (line 6). The value of each allele represents a node, therefore it has a value in a range between -1 and N-1. Where -1 is used as a special symbol to define the community boundary, whereas N represents the number of nodes in the graph. Furthermore, in this generation process, the alleles representing the community boundary are added with a probability *compb*.

Then, the population evolves a maximum number of generations (*ngen*). However, if the best μ individuals are the same during a number of generations (*nconv*) the algorithm stops, considering that the evolutionary process has converged (line 4 and 23). From the second generation, the genetic operators (crossover and mutation) are applied to create the new offspring (lines 8 to 19). Initially, two individuals are randomly selected from the n-best individuals of the current population. Then a two point crossover operation is performed over the selected parents, resulting in two new offspring. Finally, some of the new individuals are randomly chosen, using a predefined mutation probability (*mutpb*), to perform over them an uniform mutation operator. This operator replaces the value of the chosen alleles (using probability *indmutpb*) with a uniform random values selected between 0 and and N-1.

Once the new offspring population is generated, all the new individuals are evaluated using the fitness function (line 20). As described in previous subsection, this function optimizes two objectives to identify communities with high internal connectivity (m_{int}), and low external connectivity (m_{ext}) against the rest of communities found. Both measures can be chosen as parameters of the algorithm from the internal and external connectivity metrics detailed in the description of the fitness funcition implemented.

The algorithm performs an elitism selection according to the NSGA-II approach [53] that creates a Pareto-optimal Front using the two objectives functions. The best μ individuals remain from one generation to the next (lines 20 and 21). Finally, the Pareto Front of the final population is calculated to retrieve the best non dominated individuals of the evolution which are returned as final solutions (line 24).

Algorithm 6: Multi-Objective Genetic Algorithm for Overlapping Community Detection using a Node-based encoding (Node-based MOGA-OCD).

```
Input: A graph $N = (V, E)$ where $V$ represents the set of vertices denoted by
       $\{v_1, \ldots, v_n\}$, and $E$ is the set of edges $E$ denoted by $e_{ij}$ that represents a
       connection between the vertices $v_i$ and $v_j$. Parameters $m_{int}$ and $m_{ext}$ represent
       the connectivity metrics used as the optimization criterias. Positive numbers
       $ngen$, $compb$, $nconv$ $\mu$, $\lambda$, $mutpb$ and $indmutpb$ represents the main MOGA
       parameters to be fixed.
Output: HoF contains the best individuals
$C \longleftarrow \emptyset$;
$i \leftarrow 0$;
$convergence \leftarrow 0$;
while $i \leq ngen \wedge convergence = 0$ do
    if $ngen = 0$ then
        $C \leftarrow InitRamdomPop(\lambda, |V| - 1, compb)$;
    else
        $C \leftarrow Cbest$;
        for $j \leftarrow \mu$ to $\lambda$ do
            $ind1 \leftarrow RandomSel(Cbest)$;
            $ind2 \leftarrow RandomSel(Cbest)$;
            $ind1, ind2 \leftarrow Crossover(ind1, ind2)$;
            $mutchoice = Random(0, 1)$;
            if $mutchoice < mutpb$ then
                $ind1 \leftarrow Mutation(ind1, indmutpb)$;
            $mutchoice = Random(0, 1)$;
            if $mutchoice < mutpb$ then
                $ind2 \leftarrow Mutation(ind2, indmutpb)$;
            $C \leftarrow C \cup \{ind1, ind2\}$;
    $F \leftarrow Fitness(C, m_{int}, m_{ext})$;
    $Cbest \leftarrow SelNBestNSGA2(C, F, \mu)$;
    $i \leftarrow i + 1$;
    $convergence \leftarrow CheckConvergence(Cbest, nconv)$;
return $ParetoFront(C)$;
```

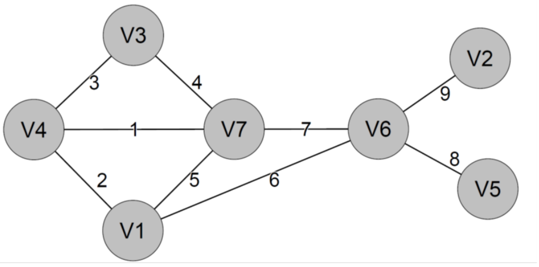
4.2 Edge-based MOGA-OCD Algorithm

The State of the Art chapter present several studies using the edge information of the graph to detect communities. Particularly, these methods are usually applied to real world networks, which tend to be sparse, and where the node-based methods often have difficulties to find large communities. For this reason, a new version of the MOGA-OCD algorithm, based on a edge-encoding, has been designed and implemented. The rest of the algorithms developed in this thesis use a node-based encoding, so the assessment of both approaches will allow to make a comparative study of these different strategies. This section presents this new algorithm, explaining in detail the edge-base encoding used, the decoding process that is required to convert this encoding to the resulting communities, and the genetic operators applied.

4.2.1 Edge-based Encoding

The edge-based MOGA-OCD Algorithm has been designed using the encoding proposed by Shi et. al. [163] and using in the algorithm called GaoCD. The authors proposed a new encoding schema based on the edges, where the number of communities is automatically determined into the evolutionary process.

In node-based GAs for CD, each allele of the chromosomes represents a node belonging to the graph. However, in this new encoding a particular allele correspond to a edge of the graph. Therefore, given a graph $G = (V, E)$, an individual of the population will have a size equal to m, where this number is the total number of edges into the graph ($m = |E|$). The position of each allele (i) correspond to a particular edge denoted by denoted by $\{e_1, \ldots, e_m\}$ that represents a connection between two vertices, and it takes a random value between the adjacent edges to the specified edge (e_i). In graph theory, two edges are adjacent if they share one node in an undirected graph. Figure 4.2 illustrates an input graph, and a possible example of a chromosome according to this edge representation. As shown in this Figure, there are 5 adjacent edges (2,3,4,5 and 7) associated with the edge 1. Therefore, the value of the first allele corresponds to one of these edges (specifically the edge 3 in this case).



3	5	2	1	4	7	9	6	8
1	2	3	4	5	6	7	8	9

Figure 4.2: Chromosome representing a graph that contains nine edges. Each allele position refers to a particular edge of the graph, and its value corresponds to a adjacent edge to this specific edge. For example, the edge number 9 has three adjacency edges (6, 7 and 8), so the value of this allele is 8 that corresponds to one of these edges.

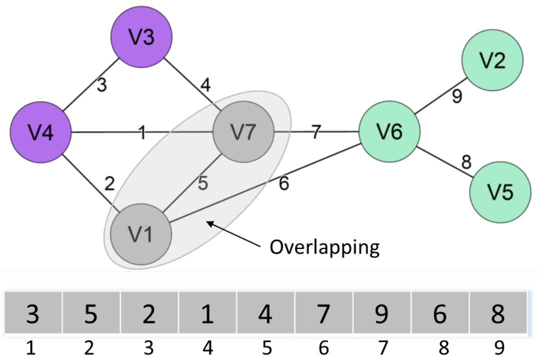
Any individual based on this encoding, cannot be directly transformed into node communities of the graph as happens with the node encodings previously used. In this case, a **decoding phase** is necessary to extract the resulting communities represented by each individual. The process carried out by this phase can be seen in Algorithm 7, and it is divided into two main parts: the process for extracting the communities of edges directly from the chromosome; and the subsequent transformation process of these communities of edges into overlapping communities of nodes.

If the value of an allele in the position i is j, this means that the edge e_i and the edge e_j have one node in common, and therefore they should be assigned to the same community. Based on this idea, the communities of edges are generated as shown in Algorithm 7 (lines 1 to 14). Then, the overlapping communities of nodes can be extracted according to the link-based algorithm proposed by Ahn et. al. [9]. Following this approach, each edge community is transformed into a node community, which contains the source and target node for each edge belonging to the original community (lines 15 to 23).

Algorithm 7: Decoding Algorithm to transform an individual with edge-based encoding into a overlapping node partition of a graph.

```
Input: A graph $N = (V, E)$ where $V$ represents the set of vertices denoted by
       $\{v_1, \ldots, v_n\}$, and $E$ is the set of edges denoted by $\{e_1, \ldots, e_m\}$ that represents a
       connection between the vertices. An individual $I$ that is a set of alleles denoted
       by $\{i_1, \ldots, i_m\}$ where each $i_k$ represents an edge $e_k$, and $m$ is the total number of
       edges ($m = |E|$).
Output: commNodes containing the communities of nodes decode from the individual
1  $commsEdges \longleftarrow \emptyset$;
2  $commE \longleftarrow \emptyset$;
3  $edgeUsed \longleftarrow \emptyset$;
4  while $|edgeUsed| < |E|$ do
5      $edgeid \longleftarrow GetAndRemoveFirstElement(I)$;
6      while $edgeid \notin edgeUsed$ do
7          $edgeUsed \leftarrow edgeUsed \cup \{edgeid\}$;
8          $commE \leftarrow commE \cup \{edgeid\}$;
9          $edgeid \longleftarrow GetAndRemoveElementAt(I, edgeid)$;
10     if $edgeid \notin commE$ then
11         $commsEdges \leftarrow commsEdges \cup commE$;
12     else
13         $comm \leftarrow GetCommWith(commsEdges, edgeid)$;
14         $commE \leftarrow commE \cup comm$;
15 $commsNodes \longleftarrow \emptyset$;
16 for $i \leftarrow 1$ to $|commsEdges|$ do
17     $commE \longleftarrow commsEdges_i$;
18     $commN \longleftarrow \emptyset$;
19     for $j \leftarrow 1$ to $|commE|$ do
20         $idNodeSource \leftarrow GetIdNodeSource(G, commE_j)$;
21         $idNodeTarget \leftarrow GetIdNodeTarget(G, commE_j)$;
22         $commN \leftarrow commN \cup \{idNodeSource, idNodeTarget\}$;
23     $commsNodes \leftarrow commsNodes \cup commN$;
24 return $commsNodes$;
```

Figure 4.3 illustrates an example of the final result obtained after applying the decoding algorithm to an individual based on a edge encoding. In this example, the individual encodes a graph partition into two communities with two overlapping nodes between them. The resulting edge communities of the first part in the decoding process will be $\{3, 2, 5, 4, 1\}$ and $\{7, 9, 8, 6\}$. Finally, applying the second part of the process, these edge communities are transformed into node communities by the extraction of the nodes associated to each edge. For example, the edge 3 in the first community gives rise to the membership of nodes V3 and V4 in the corresponding community of nodes. Therefore, continuing with the example, the node communities corresponding to the previous edge communities mentioned will be $\{V3, V4, V1, V7\}$ and $\{V6, V7, V2, V5, V1\}$.



1	2	3	4	5	6	7	8	9
3	5	2	1	4	7	9	6	8

Figure 4.3: Example of an individual with an edge-based encoding, and its corresponding node communities after performing the decoding process. The input graph has 9 edges, so the size of the individual is according to this value, and there are two resulting communities containing an overlap of two nodes (V1 and V7).

4.2.2 Algorithm Description

In this new edge-based version of the algorithm, the population evolves according to a Multi-Objective Genetic Algorithm where the individuals are evaluated using a POF, in the same way as the node-based version of the algorithm. In addition the optimization criteria used as objective functions are also the same described for this algorithm. Therefore, the pseudo-code corresponding to both algorithms is the same as shown in Algorithm 6, except for the application of the next three steps:

1. The first population is randomly generated (*InitRandomPop* method in line 6), taking into account that in the edge-based encoding, each allele represents a particular edge of the graph, and it has to get a random value only between the adjacent edges to this specified edge.

2. The mutation operator has been modified to guarantee that new generated individuals satisfy the encoding rules (*Mutation* method in line 18). Applying this operator, some of the new individuals are randomly selected using a predefined mutation probability (*mutpb*). Using the selected individuals, the mutation operation is performed, where some chosen positions of the individual (using probability *indmutpb*) can be replaced with a random value corresponding to a adjacent edge. For the crossover operator no changes are needed due to its application fulfil the encoding rules. This operator exchange alleles from two individuals randomly selected, and whose values are valid in both cases, so the new individuals created are valid too.

3. To compute the connectivity metrics used in the fitness function (*Fitness* method in line 20), it will be necessary to have a partition of the graph into communities of nodes. So, each individual should be decode before computing these metrics following the procedure described in Algorithm 7.

4.3 Experimental Results

In order to evaluate the performance of these two new evolutionary algorithms (node-based and edgde-based MOGA-OCD) to discover overlapping communities, three main phases have been carried out. The first one is focused on the analysis of the different measures, related to the network connectivity, which later will be used as optimization criteria for both multi-objective algorithms. For this purpose, a comparative assessment of internal and external metrics previously described is performed using a well-know dataset. This analysis will allow us to identify which are the most appropriate metrics for detecting overlapping communities, and to tune up both versions of the MOGA-OCD algorithm.

Then, in the second phase, the new algorithms have been compared with the results obtained using the previous evolutionary approach (GCF-II algorithm) in order to analyse the differences of results obtained by single and multi-objective approaches. Finally, in the third phase, both algorithms are assessed against different community detection algorithms extracted from the state of the art. Two measures have been computed to evaluate the effectiveness of the different algorithms; the Modularity and the Normalized Mutual Information (NMI). All the algorithms used for the experiments identify overlapping communities, therefore, the Modularity and NMI calculated are extended versions that can be applied to overlapping partitions. The following subsections describe the datasets used for the experiments, and show the analysis of the experimental results obtained for each phase.

4.3.1 Dataset Description and Experimental Setup

In the experimental phase, the MOGA-OCD algorithm has been evaluated measuring its quality and accuracy. For this purpose, a dataset collection of real world networks has been selected taking into account two main criteria: networks providing the ground truth of their community structure, with their labels to measure the accuracy; and public and widely used networks to compare the algorithm against others community detection algorithms. Table 4.1 shows the details of the eight selected networks used for the experimental analysis. This table shows for each network its total number of nodes (Nod.) and edges (Edg.), the number of communities (Com.) contained within it, and if such network includes its ground truth information (GT). In addition, two global measures of the network topology are shown (GCC and CN) to provide an overview of the structure of the network regarding to its transitivity, and the size of the cliques contained in it.

On the other hand, Table 4.2 shows the set-up for both versions of MOGA-OCD algorithm that has been used during the experimentation phases. The different algorithm parameters have been tuned up experimentally. The algorithm has been executed using a specific dataset (karate) whose ground truth of the community structure is known to measure the accuracy achieved. The best results obtained for this dataset has been used to fix the algorithm parameters.

4.3.2 Analysing the feasibility of the Connectivity Network Metrics as objective functions

As previously mentioned, the fitness function implemented optimizes two objectives. One is based on internal connectivity metrics to find communities with high internal connections be-

Dataset	Description	Nod.	Edg.	GCC	CN	TPR	GT	Com.	Ref.
nba_schedule	Games played in the 2013-2014 NBA season	30	421	0,99	15	1	Yes	6	[171]
southernwomen	Southern women social groups	32	93	0	2	0	No	-	[52]
karate	Zachary's karate club	34	78	0,26	5	0,94	Yes	2	[129]
senate_voting	Senate voting data in 2014	87	1803	0,97	45	1	Yes	3	[84]
football	American football games in 2000	115	613	0,41	9	1	Yes	12	[129]
revolution	Colonial American dissidents	261	319	0	2	0	No	-	[67]
pgp	Interactions in pretty good privacy	10680	24340	0,37	25	0,44	No	-	[12]

Table 4.1: Description of the dataset collection of real networks used for the experimental phase.

Parameter	Description	Node-based	Edge-based
λ	Population size	1500	1000
compb	Probability for each allele to be a boundary community	0,1	-
μ	Number of individuals to select for the next generation	200	100
ngen	Maximum number of generations	100	50
nconv	Number of generations with same POF to stop	10	10
mutpb	Probability of mutating an individual	0,1	0,1
indmutpb	Independent probability for each allele to be mutated	0,1	0,1

Table 4.2: Genetic Parameters of MOGA-OCD algorithm.

tween their nodes. And the other one, is based on external connectivity metrics, to detect communities having sparse externally connections to the rest of nodes of the graph. Therefore, to select the most appropriate metrics to detect overlapping communities, a comparative assessment of the connectivity metrics (internal and external) is carried out for both versions of MOGA-OCD algorithm (node-based and edge-based). In these experiments the karate network has been used as a training dataset (see Table 4.1 for details), and each test is running 10 times using a different combinations of connectivity metrics as objective functions. Finally, the best solution for each execution is selected as the final result. The following sections presents the results obtained for each version of the algorithm with its corresponding analysis.

4.3.2.1 Results for Node-based MOGA-OCD algorithm

The results obtained by the MOGA-OCD algorithm, using a node-based encoding for each combination of a pair of connectivity metrics (internal and external), is shown in Table 4.3. Since each test combining a pair of metrics is running 10 times, this table presents the average and standard deviation values obtained by the different connectivity metrics.

Firstly, the results for each external metric used as optimization criteria (m_{ext}) have been analysed. The chosen dataset contains two communities, so the Separability(Sep) and Expansion (Exp) metrics obtain the best results regarding to the number of detected communities, getting the worst results for the Cut Ratio (CR) metric. Analysing the results given by the connectivity metrics, the design of the fitness function for MOGA-OCD algorithm should be considered. This function has two objectives to identify communities with high internal connectivity (m_{int}), and sparse external connectivity (m_{ext}). According to the metric definitions presented in the State of the Art, this means that all internal metrics which are Density (D_{avg}), Triangle Participation Ratio (TPR_{avg}), Clique Number (CN_{avg}) and Clustering Coefficients (LCC_{avg} and GCC_{avg}), should be as high as possible. However, the external metrics should be as high as possible for the Separability metric, and as low as possible for the others two metrics (Exp_{avg} and CR_{avg}). Taking this into account, Table 4.3 shows that the best results for almost all the connectivity metrics have been obtained using the Sep as m_{ext} optimization parameter of the algorithm. For this specific case (first row of Table 4.3), all the internal metrics and the Separability metric achieves the higher values. In addition, the Expansion and CR metrics reach the lower values, meaning that the external connection are minimized. Therefore, this metric has been selected as one of the objective functions of the algorithm (m_{ext}).

Metric		**N[C/Nod]**	**D**$_{avg}$	**LCC**$_{avg}$	**TPR**$_{avg}$	**GCC**$_{avg}$	**CN**$_{avg}$	**Exp**$_{avg}$	**Sep**$_{avg}$	**CR**$_{avg}$
m_{ext}	m_{int}									
Sep	LCC	[2/13,6]	0,32±0,07	**0,76**±0,10	**0,95**±0,05	0,48±0,12	4,44±0,30	1,48±0,79	1,78±0,67	0,070±0,025
	GCC	[2/14,8]	0,26±0,05	0,57±0,10	0,74±0,10	**0,55**±0,09	**4,9**±0,21	1,35±0,21	1,46±0,24	0,070±0,10
	TPR	[5,9/6,3]	0,44±0,13	0,36±0,34	0,47±0,42	0,25±0,23	2,95±0,96	2,22±1,09	0,91±1,51	0,079±0,031
	CN	[2/16,6]	0,30±0,04	0,64±0,11	0,92±0,06	0,40±0,06	4,8±0,35	**0,64**±0,13	**4,12**±1,78	**0,039**±0,007
	D	[2,7/11,8]	**0,47**±0,20	0,63±0,14	0,82±0,14	0,54±0,12	4,4±0,84	2,07±1,40	1,50±0,84	0,085±0,038
Exp	LCC	[2/13,5]	0,34±0,12	0,73±0,10	0,92±0,06	0,44±0,09	4,35±0,47	1,03±0,20	1,96±0,54	0,050±0,006
	GCC	[2,1/13,9]	0,25±0,09	0,56±0,15	0,72±0,21	**0,55**±0,14	4,25±0,68	1,16±0,39	1,57±0,65	0,059±0,020
	TPR	[10,4/2,6]	0,03±0,09	0±0	0±0	0±0	1,1±0,32	1,92±0,36	0,01±0,02	0,063±0,009
	CN	[2/14,3]	0,29±0,06	0,63±0,10	0,87±0,09	0,40±0,05	4,4±0,52	0,79±0,26	2,63±1,50	0,042±0,012
	D	[2,6/11,6]	0,37±0,28	0,53±0,26	0,76±0,22	0,40±0,28	3,5±0,82	1,68±0,82	1,07±0,46	0,074±0,021
CR	LCC	[2,1/10,8]	0,34±0,08	0,74±0,13	0,89±0,13	0,42±0,14	3,78±0,44	1,47±0,47	1,27±0,68	0,063±0,015
	GCC	[4,2/7,4]	0,19±0,10	0,37±0,18	0,39±0,19	0,54±0,32	2,9±0,32	1,44±0,49	0,43±0,21	0,056±0,025
	TPR	[11,9/2,6]	0,21±0,36	0±0	0±0	0±0	1±0	2,08±0,58	0±0	0,066±0,017
	CN	[3,3/8,4]	0,32±0,08	0,39±0,19	0,57±0,18	0,43±0,11	3,45±0,72	0,94±0,22	1,33±0,79	0,037±0,10
	D	[7,6/5,8]	0,42±0,25	0,29±0,23	0,39±0,31	0,32±0,22	2,7±0,45	1,64±0,42	0,45±0,11	0,059±0,021

Table 4.3: Comparative assessment of Connectivity Network Metrics using the karate dataset for the node-based MOGA-OCD algorithm. The best average value reached by each metric is highlighted in bold. Internal connectivity metrics are marked in light grey, whereas External connectivity metrics are marked in dark Grey

Using the Sep as one of the objective functions (m_{ext} parameter), Table 4.3 shows that the best values for all the external metrics are reached using the CN as second objective function. On the other hand, many of the internal metrics (D_{avg}, LCC_{avg} and GCC_{avg}) obtain the best results when themselves are used as second objective function (m_{int}). This is an expected result, because any multi-objective algorithm try to find solutions optimizing simultaneously both objectives (in this case the external and internal connectivity metrics). However, LCC and

GCC also achieve the best results for TPR_{avg} and CN_{avg}, respectively.

It can be appreciated that there is not only a specific metric which achieves the best results. Therefore, in order to select the most suitable metric to optimize the internal connectivity, it is necessary to carry out a more detailed analysis. For this purpose, Figure 4.4 shows the results grouped by internal metrics, showing the standard deviation to study the stability of the solutions found, too. As shown in this Figure, using the TPR metric, the algorithm achieves the worst results, with the lowest values for almost all of the metrics. In addition, analysing the standard deviation, this metric obtains high values, which means that the solutions detected are quite variable between different algorithm executions. D also reaches high values for the standard deviation. For this reason, although this metric obtains good average values for the rest of the metrics, it is not being the most suitable to solve the problem. On the other hand, GCC, LCC, and CN achieve high or medium values for all the internal metrics. In addition, these metrics have lowest values regarding to the standard deviation, which means that the solutions optimizing these metrics are more stable.

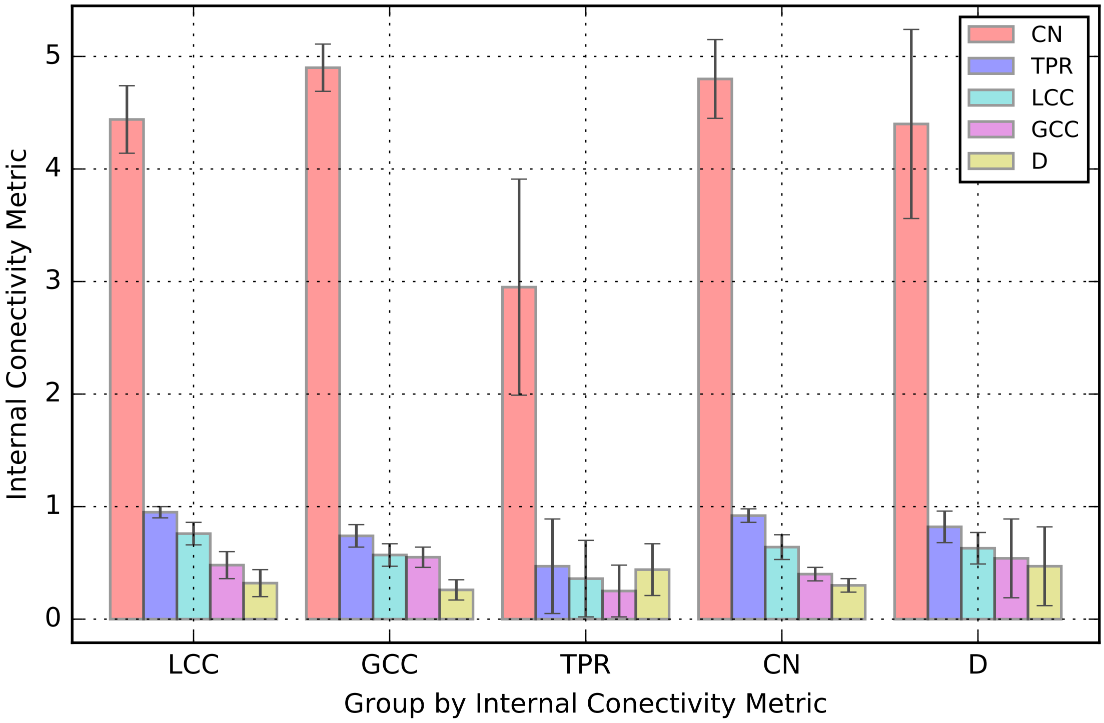


Figure 4.4: Results of network metrics grouped by internal connectivity metrics, using Sep as external function objective, in the node-based MOGA-OCD algorithm.

Finally, and in order to evaluate the quality of the results for each internal connectivity metric, the Modularity and NMI measures have been calculated and shown in Table 4.4 and Figure 4.5. According to the results presented in previous table, in general terms the TPR obtains the worst results, reaching bad outcomes on both metrics. It can be clearly noticed that the CN obtains the best values regarding to the Modularity, but contrary to the NMI this metric gets the worst values. The GCC achieves the greatest precision, finding the solutions most similar to the ground truth for the karate dataset. However, the value of the Modularity is low. As show in Figure 4.5, taking into account both evaluation measures, the LCC obtain good overall values, reaching a value very close to the best result of NMI and quite high related to the Modularity.

In general terms, LCC achieves good results related to the rest of internal metrics, and the best results according to the quality of the detected solutions. So, this metric has been selected as the second objective function of the algorithm (m_{int}). Based on these experimental results, we can conclude that the optimization criteria which perform better in the algorithm are the *Sep* as objective function for m_{ext} parameter, and the *LCC* as m_{int} parameter.

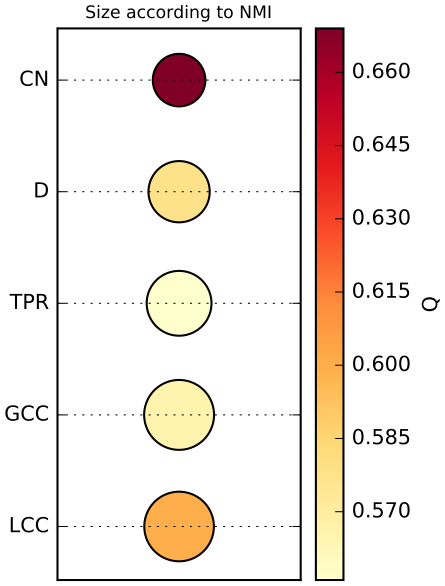


Figure 4.5: Comparison between evaluation metrics, NMI and Q, for the node-based MOGA-OCD algorithm. The size of the bubble represents the NMI, whereas the color represents the Q.

Met.	**NMI**$_b$	**NMI**$_{avg}$	**Q**$_b$	**Q**$_{avg}$
CN	0,129	0,094±0,024	**0,669**	**0,373**±0,081
D	0,175	0,137±0,020	0,578	0,261±0,131
TPR	0,196	**0,169**±0,017	0,556	0,146±0,114
GCC	**0,228**	0,144±0,035	0,566	0,244±0,119
LCC	0,226	0,152±0,039	0,600	0,316±0,105

Table 4.4: Results of evaluation metrics (NMI and Q) for each internal connectivity metric, using Sep as one of the objective functions (m_{ext}) in the node-based MOGA-OCD algorithm.

4.3.2.2 Results for Edge-based MOGA-OCD algorithm

As shown in the previous version of MOGA-OCD algorithm, a detailed study for each combination of a pair of connectivity metrics (internal and external) has been carried out using the edge-based version of the algorithm. This analysis enables to establish which of the metrics optimize better the search of communities, and therefore, these will be use as objective functions in the algorithm. Table 4.5 shows the average, and standard deviation values, obtained by the different connectivity metrics for each combination of metrics used as objective functions.

Analysing the results obtained for each external metric, when it is used as optimization criteria (m_{ext}), it can be noticed that Sep achieves the worst values in general (first row in the table). As would be expected, the higher value in relation to Sep_{avg} has been reached in this case, besides with rather difference regarding the rest of external metrics (Exp and CR). This fact also happens when the others external metrics are used as optimization criteria, but with a lower difference on the improvement. However, the results of the internal metrics are worse

Metric		N[C/Nod]	**D**$_{avg}$	**LCC**$_{avg}$	**TPR**$_{avg}$	**GCC**$_{avg}$	**CN**$_{avg}$	**Exp**$_{avg}$	**Sep**$_{avg}$	**CR**$_{avg}$
m_{ext}	m_{int}									
Sep	LCC	[2,2/20,7]	0,26±0,05	0,72±0,05	0,96±0,02	0,36±0,07	4,7±0,26	0,74±0,15	14,58±5,42	0,054±0,009
	GCC	[2/20,6]	0,31±0,17	0,72±0,06	0,96±0,01	**0,42**±0,14	**4,5**±0,53	2,37±1,92	16,46±3,93	0,102±0,065
	TPR	[2/31,45]	0,15±0	0,58±0,01	0,95±0	0,26±0	**5**±0	0,12±0,03	24,22±1,24	0,043±0,008
	CN	[2/31,4]	0,15±0,01	0,58±0,01	0,95±0	0,26±0	**5**±0	0,11±0,05	**25,56**±2,84	0,040±0,009
	D	[2,1/21,9]	0,28±0,12	0,67±0,05	0,95±0,01	0,40±0,12	**5**±0	0,67±0,30	18,83±4,27	0,046±0,008
Exp	LCC	[2,9/20,8]	0,26±0,10	0,72±0,12	**0,98**±0,02	0,36±0,08	4,7±0,35	0,89±0,45	4,19±2,03	0,057±0,016
	GCC	[2/25,5]	0,20±0,03	0,61±0,04	0,95±0,01	0,32±0,03	**5**±0	0,34±0,11	8,26±3,13	0,043±0,010
	TPR	[2/32]	0,15±0	0,58±0,01	0,95±0	0,26±0	**5**±0	**0,09**±0,02	20,67±3,93	0,049±0,009
	CN	[2/31,7]	0,15±0,01	0,58±0,01	0,96±0	0,26±0	**5**±0	0,10±0,03	22,13±6,53	0,047±0,010
	D	[3,1/18,6]	**0,37**±0,24	0,65±0,11	0,95±0,02	0,34±0,12	4,1±0,88	0,66±0,49	8,19±5,54	0,053±0,003
CR	LCC	[3,5/15,5]	0,31±0,06	**0,77**±0,08	**0,98**±0,02	0,38±0,03	4,6±0,31	0,71±0,40	7,20±2,13	0,046±0,013
	GCC	[2,4/19,2]	0,28±0,13	0,65±0,06	0,95±0,01	0,38±0,13	4,9±0,32	0,53±0,26	6,00±1,87	0,035±0,007
	TPR	[2/22,1]	0,22±0,01	0,64±0,01	0,95±0,01	0,34±0	**5**±0	0,40±0,01	7,09±0,03	**0,032**±0,001
	CN	[2/22,2]	0,22±0,01	0,64±0,01	0,94±0,01	0,34±0	**5**±0	0,40±0,01	7,09±0,03	**0,032**±0,001
	D	[3,1/15,5]	0,35±0,17	0,64±0,03	0,94±0,01	0,35±0,02	4,4±0,97	0,69±0,35	5,01±2,36	0,038±0,007

Table 4.5: Comparative assessment of Connectivity Network Metrics using karate dataset for the edge-based MOGA-OCD algorithm. The best average value reached by each metric is highlighted in bold. Internal connectivity metrics are marked in light grey, whereas External connectivity metrics are marked in dark Grey.

than the other two external metrics (Exp and CR).

Comparing the results obtained with regard to internal metrics for Exp and CR, overall, CR obtains higher average values. But, as shown in the third row of Table 4.5, there is not only a specific internal metric reaching the best values when it is used as second optimization criteria (m_{int}). Therefore, in order to chose the most suitable metrics to use in the algorithm, it is necessary to carry out a more detailed analysis. For this purpose, the Q and NMI measures have been show in Table 4.6 and Figure 4.6, to assess the quality of the solutions found when CR is used as m_{ext} parameter of the algorithm.

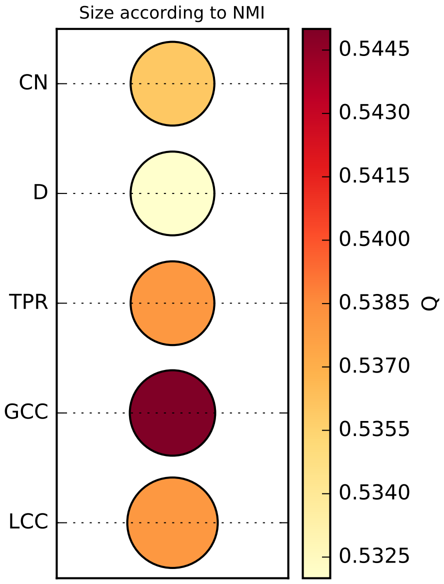


Figure 4.6: Comparison between evaluation metrics, NMI and Q, for the edge-based MOGA-OCD algorithm. The size of the bubble represents the NMI, whereas the color represents the Q.

Met.	NMI$_b$	NMI$_{avg}$	Q$_b$	Q$_{avg}$
CN	0,150	0,122±0,036	0,536	**0,534**±0,003
D	0,150	0,142±0,009	0,532	0,513±0,022
TPR	0,150	0,129±0,034	0,538	0,533±0,003
GCC	0,155	0,144±0,009	**0,545**	0,519±0,021
LCC	**0,175**	**0,157**±0,008	0,538	0,449±0,068

Table 4.6: Results of evaluation metrics (NMI and Q) for each internal connectivity metric, using CR as one of the objective functions (m_{ext}) in the edge-based MOGA-OCD algorithm.

Analysing the results shown in previous table and figure, the worst outcomes for both measures (Q and NMI) has been reached when the D is used to optimize the search for communities. According to the Q value, when the GCC is used as optimization criteria, the algorithm achieves the highest value, and an intermediate NMI value. Conversely, the algorithm is able to detect the solution most similar to the ground truth of the karate dataset, using the LCC. So this metric reaches the higher precision, and a good result for the Q value. However, analysing the results shown in Table 4.5 related to the different connectivity metrics, in general it can be noticed than LCC obtains better results than GCC for most of these metrics, reaching even the best values in some cases (LCC_{avg} and TPR_{avg}). Therefore, taking these experimental results into account, the two metrics selected as objective functions for the edge-based MOGA-OCD algorithm have been, the CR metric for m_{ext} parameter, and the LCC metric for m_{int} parameter.

4.3.3 Preliminary comparison of MOGA-OCD algorithms against previous evolutionary approach

In this section, the results obtained using the previous evolutionary approach (GCF-II algorithm) have been compared against the results of the new multi-objective approaches (node-based and edge-based MOGA-OCD algorithm). As presented in previous chapter, the GCF-II algorithm was evaluated using a dataset collection, which consists of a series of graphs representing the cast votes between the participant countries in the Eurovision Song Contest. Therefore, in order to carry out a comparative analysis with it, the same dataset collection has been chosen. The definition of the Q metric extended by Nicosia et al. [134] for overlapping partitions into graphs is used to evaluate the outcomes obtained by each algorithm.

The three evolutionary approaches are based on GAs, using as fitness function metrics related to the Graph Theory. The main difference between the previous algorithm, and the two new algorithms presented in this chapter, is that MOGA-OCD algorithms optimize two objective functions. As shown in Figure 4.7, both new algorithms significantly improve the results in all of the cases.

Since 2004, the televoting system was established to vote in the final round of the contest. From this point on, all the countries that are participant of the semifinals rounds, can also cast votes in the final round. For these reasons, the network topology of the graph representing the subsequent year changed. The graphs for the televoting period contain more nodes and edges. This fact can be appreciated in the results that have a clearly variation, as shown in the Figure 4.7. However, the results of MOGA-OCD algorithms are still much better despite this fact. Therefore, in general terms, it can be concluded that the multi-objective approach is able to perform better, than the previous one based on a single-objective of the algorithm implemented in this thesis.

On the other hand, comparing only the quality of the solutions found by MOGA-OCD algorithms, it can be noticed that the variation for the televoting period in the Q values is different in both cases. The Q values obtained by Node-based MOGA-OCD decrease in value at this period, whereas the Q values increase for the edge-based version of the algorithm. This may suggest that for graphs with higher connectivity, the edge-based algorithm outperforms the node-based version. During the following the experiments, this aspect of the MOGA-OCD algorithms will be considered more in detail to validate if this occurs with other kinds of graphs.

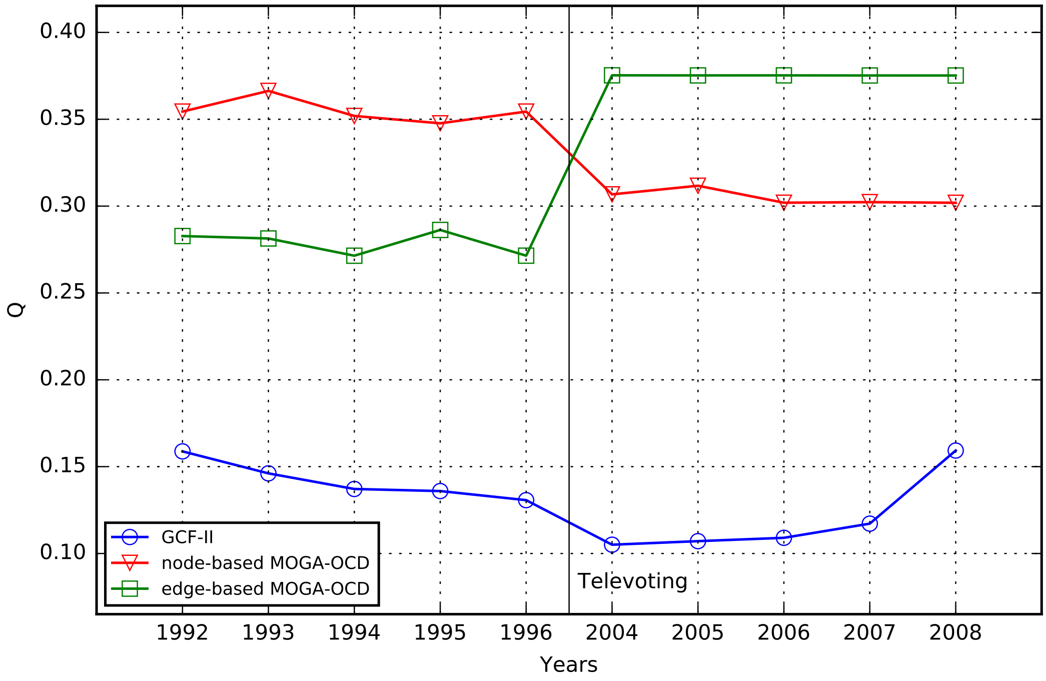


Figure 4.7: Comparative assessment of evolutionary approaches for overlapping community detection using the Q value as evaluation metric. Each year is related to a graph representing the voting of this year in the Eurovision Song Contest.

4.3.4 Comparative assessment of communities detection algorithms

This section shows an additional analysis, to evaluate the accuracy and quality of the new MOGA-OCD algorithms, against other similar algorithms from the state of the art such as Clique Percolation, Coda, Conga and Congo. For this purpose, two type of datasets have been chosen: real world networks providing the ground truth of their communities which will be used to evaluate the accuracy of these algorithms; and more complex real networks, without ground truth, which will be used to evaluate the effectiveness and quality of detected solutions.

Taking into account the conclusions extracted from the previous analysis on connectivity metrics, in the next experiments the node-base MOGA-OCD algorithm will be executed using the Sep metric, and the LCC as objective functions, whereas the CR and LCC metrics will be used for the edge-based version of the algorithm. In this experimental phase, the algorithm is executed 10 times for each dataset, and the best solution of these 10 executions has been selected as the final result to compare against the solutions generated by the others algorithms.

4.3.4.1 Experimental results for datasets with ground truth

Firstly, the different algorithms have been applied to the dataset with ground truth, measuring their accuracy and quality through the NMI and Q measures, respectively. Table 4.7 presents the results grouped by dataset, showing both metrics for each algorithm. In addition, this table shows the number of communities detected (N. Com.), and the average number of nodes per community (Avg. Nodes), in order to compare the different graph partitions performed by each algorithm.

DataSet	Algorithm	N. Com.	Avg. Nodes	Q	NMI
nba_schedule	groundtruth	6	5	0,281	1
	CPM	2	15	0,879	0,21
	Coda	20	5	0,069	0,22
	Conga	2	15	0,879	0,21
	Congo	2	15	0,879	0,21
	Node-based MOGA-OCD	2	15	0,879	0,21
	Edge-based MOGA-OCD	2	15	0,879	0,21
senate_voting	groundtruth	3	42	0,810	1
	CPM	1	87	0	0
	Coda	79	15	0	0,43
	Conga	2	44,5	0,852	0,75
	Congo	2	44,5	0,852	0,75
	Node-based MOGA-OCD	2	39,5	0,754	0,64
	Edge-based MOGA-OCD	2	44,5	0,852	0,75
karate	groundtruth	2	17	0,261	1
	CPM	3	6	0,515	0,04
	Coda	36	4	0	0,16
	Conga	4	8,5	0,476	0,11
	Congo	3	6	0,322	0,13
	Node-based MOGA-OCD	2	14,5	0,605	0,23
	Edge-based MOGA-OCD	4	10	0,443	0,17
football	groundtruth	12	10	0,642	1
	CPM	4	13	0,445	0,25
	Coda	76	7	0	0,42
	Conga	6	31,5	0,511	0,22
	Congo	11	9	0,342	0,20
	Node-based MOGA-OCD	5	9	0,183	0,48
	Edge-based MOGA-OCD	8	24,75	0,386	0,47

Table 4.7: Comparative Assessment of Community Detection Algorithms for all the datasets **with** ground truth.

As shown in this Table, most of the algorithms perform the same partition of the graph related to the first dataset (*nba_schedule*), whit the exception of Coda algorithm. This graph only contains two connected components with 15 nodes (see a further description in Table 4.1), and for this reason the algorithms tend to partition the graph into two communities, achieving the same results in terms of NMI and Modularity. However, Coda algorithm detects a greater number of communities, and it slightly improves the NMI metrics, but the Q value is much worse being close to 0, which corresponds to a modularity for a randomly generated graph.

Something similar happens using the *senate_voting* dataset. There are two highly connected components in the graph, so four of the algorithms (Conga, Congo and both MOGA-OCD algorithms) perform a partition into communities according to these components. In addition, in this case these components of the graph are very similar to the two communities given in the ground truth. Therefore these algorithms achieve high values of NMI, while the other two algorithms (CPM and Coda) obtains low values on both evaluation metrics. The best results are obtained by Conga and Congo and edge-base MOGA-OCD algorithm, showing the node-based version of the new algorithm results very close to them.

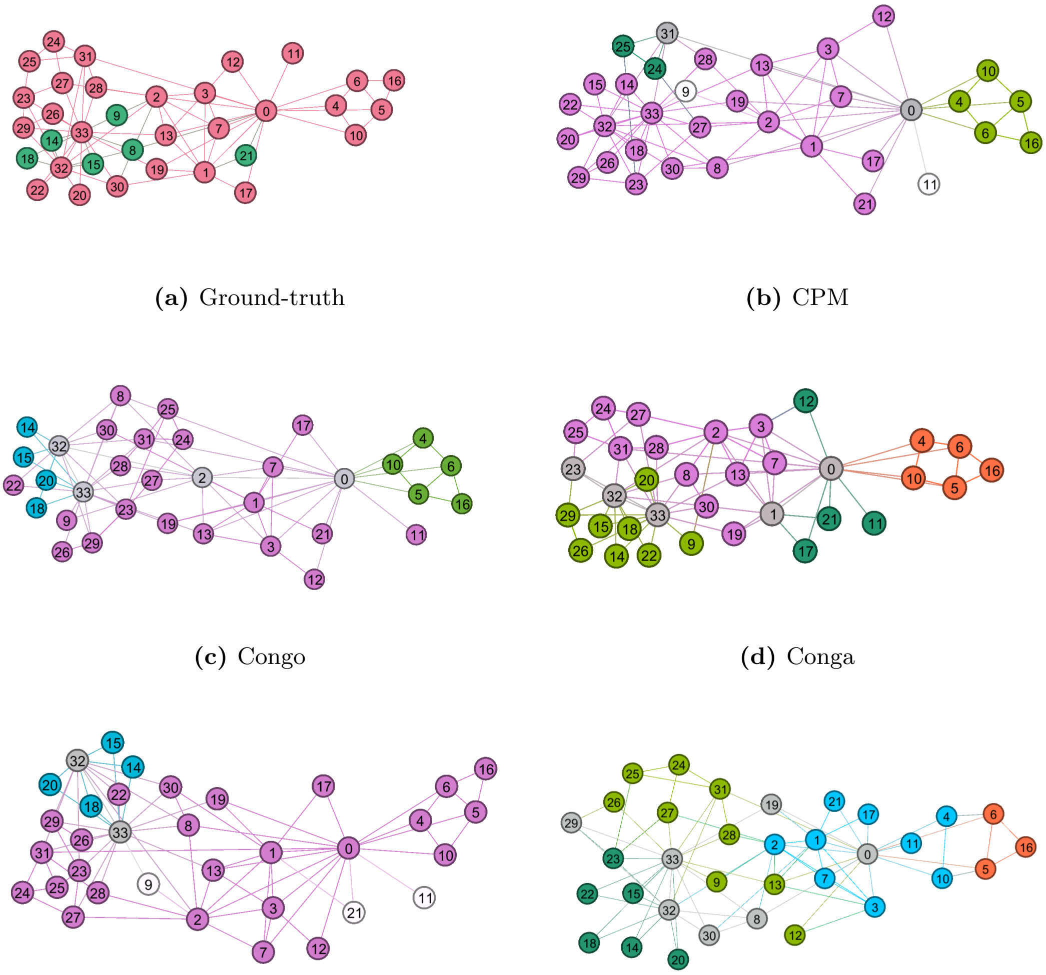
The outcomes of the following two datasets are more varied for each different algorithm, because these graphs don't have a network topology as structured as the previous graphs. Analysing these results, in the karate dataset, both MOGA-OCD algorithms outperform the NMI value, and in addition the node-based version the Q value is significantly improved. In the results for the rest of the algorithms, it can be noticed that when a specific algorithm improves a particular measure, its value in the other measure worsens. For example, the CPM algorithm has a high value on Modularity, whereas it has the worst value on NMI. By contrast, Coda algorithm obtains a higher value on NMI, having the worst value on Modularity. Only the MOGA-OCD algorithms are be able to achieve a balanced value in both measures.

To analyse more in detail the results, a visualization of the communities detected has been generated, proving a better appreciation of the structure of these node groupings. In Figure 4.8 the communities detected, using the karate dataset for each algorithm, are plotted using a graph representation. The first sub-figure (a) shows the existing communities in the ground truth of this dataset. For the rest of figures, the overlapping nodes have been coloured in gray, while the unassigned nodes are shown in white. Due to the poor outcome obtained by Coda algorithm, these results has not been included in the figure.

CPM, Congo, and Conga algorithms (Figure4.8.b, Figure4.8.c and Figure4.8.d) detect an isolated community consisting of 5 nodes, which is clearly identified according to the topology structure of the network. However, this community is not found in the ground truth related to this dataset. This is because the ground truth has been manually labelling by a human, and it may not necessarily correspond to the network structure. The edge-based MOGA-OCD algorithm splits into two this isolated community, whereas the node-based version contains this in a larger community as in the Ground-truth case. For that reason, the node-based MOGA-OCD algorithm reaches the higher NMI value. The rest of nodes belonging to the graph have been divided into 2 or 3 communities. In general terms, the resulting communities obtained by the MOGA-OCD algorithms are the most similar to the ground-truth, and also obtain a medium value of Q, meaning that are well structured too.

Finally, analysing the results of the last dataset with ground truth (football), Conga achieves the best Q value, but with a low NMI value. On the other hand, both MOGA-OCD algorithms obtain the best NMI values, but the node-based version has a low Q value, whereas the edge-based version obtains a good value on Q. These results seems similar to those obtained with the previous dataset, where for the algorithms that reach high values on Q, their accuracy considerably decrease. This fact can be clearly observed in the case of Coda algorithm that reaches the highest value of NMI for the algorithms from the state of the art, but it gets by far the worst value of Q. There has been a slight decrease in the Q value for the edge-based MOGA-OCD algorithm, but it is not close to 0 as occurs in the Coda algorithm, which means that the partition detected is still structured from the point of view of the network topology.

Taking into account the results from all datasets, it can be concluded that most of the community detection algorithms obtains similar results when they are applied to graphs with a strong structured topology. In these cases, the new MOGA-OCD algorithms identify commu-



(a) Ground-truth

(b) CPM

(c) Congo

(d) Conga

(e) Node-based MogaOCD

(f) Edge-based MogaOCD

Figure 4.8: Visualization of resulting communities from different algorithms for the karate dataset. Overlapping nodes are represented in gray colour, whereas the unassigned nodes are shown in white colour.

nities according to the network topology, with good values in both evaluation metrics. On the other hand, using datasets corresponding to less structured graphs, MOGA-OCD algorithms improve the value for NMI measure, and they reach good values for the Q measure. It means that the new evolutionary algorithms perform a good partition of the graph with regarding to the Q value, being also the most similar partition to the ground truth, so reaching overall the best results.

4.3.4.2 Experimental results for datasets without ground truth

In order to assess the effectiveness of the new algorithms, it is necessary to test them using more complex networks as shown in Table 4.8. The first two datasets, southernwomen and revolution, correspond to unstructured graphs. The CC and TPR in both datasets is 0, and they have a very low value of CN (2) too, as can be seen in Table 4.1 where a overall description of each dataset was given. These datasets are used to test the algorithm behaviour on these type of graphs. Due to the CC and CN values, the LCC metric can not be used as objective function by the algorithms. Therefore, in these two cases the chosen metric for m_{int} parameter has been the D. Table 4.8 shows the results obtained on Q for each algorithm, including also the number of communities detected (N. Com.), and the average number of nodes per community (Avg. Nodes).

DataSet	Algorithm	N. Com.	Avg. Nodes	Q
southernwomen	CPM	-	-	-
	Coda	51	4	0
	Conga	4	11	0,475
	Congo	3	14	0,665
	Node-based MOGA-OCD	2	15,5	0,664
	Edge-based MOGA-OCD	2	19	0,639
revolution	CPM	-		-
	Coda	45	57	0
	Conga	60	5	0,001
	Congo	63	5	0,001
	Node-based MOGA-OCD	4	44	0,297
	Edge-based MOGA-OCD	3	95	0,830
pgp	CPM	734	3	0,568
	Coda	100	135	0,560
	Conga	-	-	-
	Congo	-	-	-
	Node-based MOGA-OCD	4	406,5	0,124
	Edge-based MOGA-OCD	2135	8,9	0,188

Table 4.8: Comparative assessment of Community Detection Algorithms for a dataset collection **without** ground truth.

Analysing the results of the first dataset in Table 4.8, it can be seen that Congo and both MOGA-OCD algorithms achieve the best values of Q, being them values very close. However the partition of the graph performed by each algorithm is different. Congo detects three communities, whereas MOGA-OCD algorithms identify only two communities, although the average size of the communities detected is similar for all of them. The worst results have been obtained using Coda algorithm, as already occurred using the datasets with ground truth. On the other hand, CPM is not able to partition the graph. This is due to this algorithm have a minimum clique size to find communities which is 3 by default. The average Clique Number of the communities detected for the others algorithms is 2, so CPM algorithm is unable to find any community that fulfils this requirement.

Comparing the Q values obtained for the *revolution* dataset, it is clear that MOGA-OCD algorithms improve the results of the rest of the algorithms, having all the other algorithms a very low value of Q that is close to 0. In addition, using the edge-based version, this improvement is very significant. As in the previous dataset, the same effect happens with CPM algorithm that it is unable to detect any community. Regarding the number and size of communities, the new algorithms find a less number of communities, but with a larger size. However, as mentioned before, these communities are more structured, due to their higher value of Q.

Finally, Table 4.8 presents the results of the algorithms using a dataset of large size (pgp dataset described in Table 4.1). As shown in this table, two algorithms are not able to perform a partition of the graph (Conga and Congo). In this case Coda algorithm has a high value on the Q, thereby being an exception, due to for the rest of datasets this algorithm had always obtained the worst results. CPM algorithm obtains the best value of Q, but finding many communities of very small size. In particular, the average size of these communities is 3, so this method is returned as the graph partition the triangles contained in it. Otherwise, the node-based MOGA-OCD identities few communities with a large size into the graph, reaching a good value of Q. The edge-based version of the algorithm splits the graph into a large number of communities with a intermediate size, and its Q value is better than for the node-based version. Therefore, in general terms, MOGA-OCD algorithms are able to perform a good overlapping division of the different types of graphs, specifically when they are compared against the algorithms from the state of the art.

4.3.5 Experimental Conclusions

Taking into account all the experimental analysis carried out in this chapter, some main conclusions can be drawn. Coda algorithm obtains the worst results in most of the experiments. This algorithm tend to detects many unstructured communities with a very low value of Q, or even equal to 0 in same cases. In addition the community overlapping is so high, making impossible to differentiate the distinct communities, and having a precision very low for the dataset with ground truth.

Conga and Congo algorithms usually have good results when the graph is very structured and the algorithm can partition it according to its network topology. But when these algorithms are applied using large size graphs are not able to perform an overlapping partition. In addition, for unstructured or sparse graphs, the accuracy and quality of the partition detected by these algorithms significantly decrease. This means that they are highly dependent on topology and the structure of the input graph.

Using unstructured or sparse graphs, the node-based MOGA-OCD algorithm improves the results in most of the cases. Related to the structured graphs, this version of the algorithm obtains results similar, or closer, to the rest. On the other hand, the edge-based MOGA-OCD algorithm obtains good results for both types of graphs (structured and unstructured).

Therefore, in general terms, the new MOGA-OCD algorithms have good results for the different types of graphs analysed, including largest datasets such as *pgp* where the same algorithms are not able to compute a graph partition. Furthermore, using the dataset collection with ground-truth, the new algorithms reaches the best accuracy in most of the cases studied. Therefore, the experimental results show that these new approaches improve overall the results of the other classical approaches from the state of the art.

Chapter 5

Applications for detecting communities on Social Graph-based Information

"This world is not so kind, People trap your mind
It's so hard to find someone to admire."

- Madonna (Nobody knowns me - American Life)

This chapter presents the application of the community detection algorithms to identify opinion communities related to Public Healthcare topics in Social Networks, acquiring new collective knowledge about their behaviour, preferences, profiles, etc. For this purpose, both classical algorithms and new algorithms implemented in this thesis will be used comparing their results to evaluate them. In this particular work, these algorithms are applied to detect communities in Twitter which are disseminating vaccine opinions. Finally, an analysis of the influence of these communities to the rest of users, in a particular zone or country, is carried out proving how useful is this new knowledge to Public Healthcare Organizations. These organizations could improve their strategies increasing control and preventive measures in the risk zones identified.

For this purpose a dataset collected from Twitter, and the vaccination coverage rates retrieved from the immunization monitoring system of WHO, have been used to perform several analysis. Using both datasets, an initial analysis is made focused on measuring the potential influence of vaccine opinions based on the variation in the coverage rates. In this analysis two factors are used: Topic Relevance Factor (quantifying the relevance of vaccine topic in a given country) and Kurtosis of Vaccination Coverages (measuring the 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.

5.1 Detecting Discussion Communities on Vaccination in Twitter

The use of vaccines has contributed to dramatically decrease mortality rates from infectious diseases in the 20th century [72]. 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 [95, 136, 169]. These vaccination communities have taken advantage of social media technologies to effectively disseminate its message and to spread their theories [99]. In recent years, several studies on various social media services such as YouTube [102], MySpace blogs [101], and Social Networks (SN) [160], present this dissemination and their effects. In addition, statistical analysis show how this vaccination information influences social media users in their treatment decisions [166].

Currently, one of the most popular social networks (SN) is Twitter [4], producing huge amounts of public information. Twitter users can generate new data sources of collective intelligence through their comments and interactions, allowing the application of data mining techniques in several fields [21] such as marketing campaigns [38, 23], financial prediction [13] or public healthcare [16, 19, 45], amongst others.

In the related literature, there are several works investigating knowledge acquisition from social networks about vaccine sentiments using classification techniques [33, 113, 153] in most cases. 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 [154]. 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 section presents an application of community detection algorithms to detect user groups in Twitter which are disseminating vaccine opinions in order to analyse their influence to the rest of users into their own 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 the community detection techniques, increasing control and preventive measures in the identified risk zones. In this particular work, the use of these algorithms 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. Afterwards, a second part of the work is focused on the study of a real re-tweet graph, representing the user interactions who talk about vaccination.

5.1.1 Methodology to analysis the potential influence of social vaccine communities on healthcare

This preliminary analysis 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. 5.2). 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. 5.3). 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.

To carry out this 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) [5] have been retrieved. 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.

The whole process of influence analysis includes four different sub-phases: Data Extraction, Data Preprocessing to Identify Tweet Locations, Social Data Analysis, and Visualization of Geo-Spatial Information. These are detailed in the following subsections.

5.1.1.1 Data Extraction

The data used in this work have been extracted from two sources:

- **Twitter** [4]: 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 that 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** [5]: 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.

5.1.1.2 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 on-line 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 process has been divided into two further steps:

1. **Geocoding Process**: It is the process of converting addresses (for example: "Mountain View, California") into geographic coordinates (37.42, -122.08). For this purpose, 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 Using this library the geographic coordinates have been used to identify the origin country from the users.

5.1.1.3 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:

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. The web page of Statista [1] provides information on the Twitter penetration per country. This measure is defined as the number of active twitter users relative to the total amount of internet users. Taking into account this data, and the information about internet users by country extracted from Internet Live Stats [3] (see Figure 5.1), the TR_f factor for each particular country is calculated as follows:

$$TR_f(C) = \frac{\%NTwitterVaccineUsers}{\%TwitterPenetration \times \%InternetUsers} \tag{5.1}$$

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 same country. Finally, to scale the factor values into a range of $[0, 1]$, an unity-based normalization is applied:

$$\overline{TR_f(C)} = \frac{TR_f(C) - min(TR_f(C))}{max(TR_f(C)) - min(TR_f(C))} \tag{5.2}$$

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 correspong to a distribution with a sharper peak and fatter tails, whereas a low kurtosis distribution has a more rounded peak and thinner tails [54]. In this work, the kurtosis value is calculated according to the Fisher's definition [68] 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. According to the variation of the coverage vaccination rates, high *kurtosis* values represent a sharp change on these rates. 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 as shown below:

$$K_{VCR}(C) = n\frac{\sum_{i=1}^{n}(X_i - X_{avg})^4}{(\sum_{i=1}^{n}(X_i - X_{avg})^2)^2} - 3 \tag{5.3}$$

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.

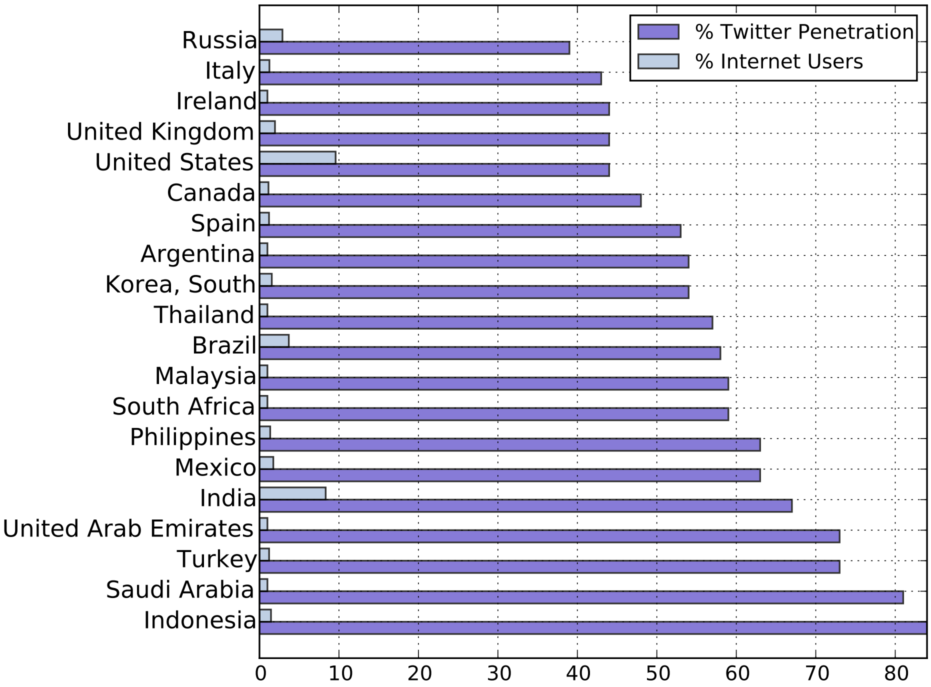


Figure 5.1: Penetration of Twitter Users and Internet Users per country (Top 20). Data taken from Statista and Internet Live Stats.

Finally, to validate the potential influence on immunization coverages of twitter opinions, the ***Spearman correlation coefficient*** [91] 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. On the contrary, 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 [185]. In our work these critical values are used to validate whether a significant correlation between both factors exits.

5.1.1.4 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 particular 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.

5.1.2 Vaccine Community Detection in Twitter

The main phase of this work is focused on the study of 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, to carry out the social analysis of this re-tweet network, two different steps are performed: Finding Social Communities and Analysing Social Network Data.

5.1.2.1 Finding Social Communities

Community detection algorithms have been studied extensively in computer science [154] but in particular, for social media mining [13, 184]. 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 in the state of the art, 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 introduced in the section correspond to the state of the art, and they are as follows: Fast-Greedy, InfoMap, Leading Eigenvector, Label Propagation, Multi-Level and Walktrap.

The evaluation of the community outcomes after applying the previous algorithm is a difficult task. This is because the list of each community members is rarely known, as in this specific case where the data used are gathered from Twitter. 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. In this work the metrics used to compare the results of all the algorithm are: Q, D, TPR, LCC and CN.

5.1.2.2 Analysing Social Network Data

After one algorithm is chosen to detect the anti-vaccine communities, there are several networks metrics that can be used to discriminate the most relevant users within each community allowing 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 [184]:

1. *Degree Centrality*: It used to analyse the user importance through their interactions. 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).

5.1.3 Experimental Results

5.1.3.1 Dataset Description

As described in previous sections, the data collected to perform the data analysis of vaccination influence was extracted from two sources: Twitter APIs, and WHO Web Site. In particular the data extracted from each source are as follows:

1. **Twitter APIs** [4]: 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 5.1 shows the number of tweets gathered during this time period. As shown in this table, less than the **1%** of all tweets are geo-located. However, after performing the preprocessing to identify tweet locations (described in previous subsection), this value increases noticeably up to **51%**.

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

Table 5.1: Number of tweets on vaccine related topics during seven months.

2. **WHO Web Site** [5]: 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.
   (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.

Using these information two factors (TR_f and K_{VCR}) have been calculated to measure the social influence of social communities in Twitter on immunization coverage. 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.

5.1.3.2 Preliminary analysis for Topic Relevance Factor (TR_f)

Figure 5.2 shows the results obtained for TR_f factor, where the values are sorted generating a ranking of countries by relevance. It can be noticed that there are only few countries with large values while most countries have very low values. Figure 5.2 shows the top five countries that 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.

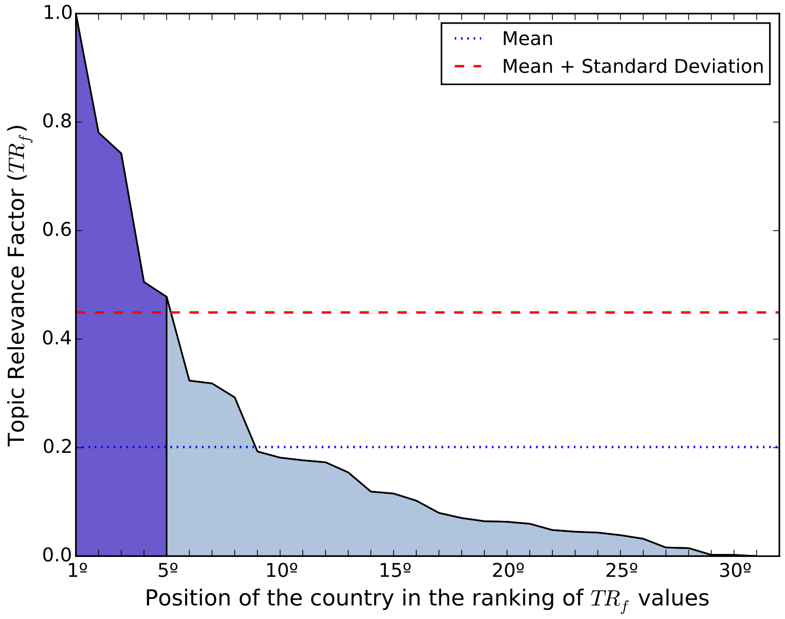


Figure 5.2: Values of TR_f for the countries talking about the vaccination topic. The results are shown in order as a ranking of countries by relevance.

To continue the analysis of results, a geo-spatial visualization has been performed to allow for a visual analysis of the social data across all countries. For this purpose, a map summarizing the location information of the users talking about vaccines is generated to identify the regions of interest. This map is generated according to the TR_f results obtained for each country as shown in Figure 5.3. Analysing the results, 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 opining 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 highly probable that the users opinions influence over vaccination coverage.

5.1.3.3 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 the most relevant, and have been selected in order to continue the analysis. Table 5.2 shows the results of both factors computed for these top five countries. If vaccine opinions affect

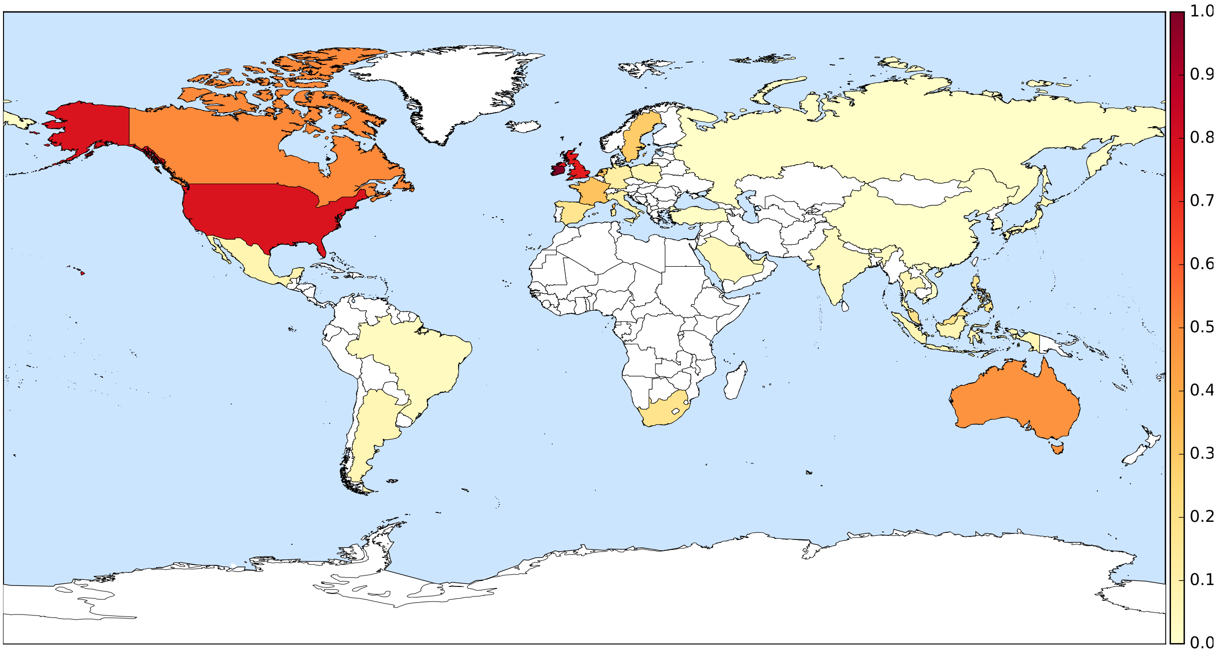


Figure 5.3: World Map based on TR_f, measuring the relevance of vaccination topics for each country.

user vaccination decision making, the immunization coverage rates of vaccination would show variations which 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 5.2, 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.

Country	***N. 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 5.2: 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.

To identify 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 purpose, the *Spearman correlation coefficient* [91] 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 countries with higher variations in vaccination rates. The results obtained can be seen in Table 5.3.

In Zar [185] study, 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 in Table 5.2, there is no data available for HepB3 vaccination coverages for United Kingdom. Therefore, the minimal critical value is 0.6 for this specific case. Analysing the results shown in Table 5.3, 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 due to there are also supporting movements trying to increase immunization rates, as in the case corresponded to MCV vaccine for this study.

Vaccine	**Spearman Coefficient**	**pValue**
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 5.3: 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.

5.1.3.4 Detecting communities on vaccination discussion

This section reports an additional analysis related to the social influence from twitter opinions 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, so these users have been selected to generate the network for the social study.

As mentioned in the state of the art, there are several community detection algorithms which can be applied to solve this problem [28, 41, 130, 146, 147]. These algorithms can be divided into two categories (Overlapping or Partitional) depending on the graph partitioning is performed allowing overlaps between the different partitions. Therefore, firstly a comparative assessment of several algorithms is carried out to choose the most appropriate for each category. In this way, the difference between the two categories of algorithms can be studied. Table 5.4 shows for each algorithm if it detects overlapping communities (Overlap), as well as the number of communities (N. Com.) identified by it, and the average number of nodes of these communities (Nodes$_{avg}$). In addition, four measures of the network topology are shown (CN$_{avg}$, D$_{avg}$, LCC$_{avg}$ and TPR$_{avg}$) to provide an overview of the structure of the communities detected regarding to its internal connectivity, and the size of the cliques contained on it. Finally, to evaluate the quality of the communities, the Q measure has been shown.

Algorithm	**Overlap**	**N. Com.**	**Nodes**$_{avg}$	**CN**$_{avg}$	**D**$_{avg}$	**LCC**$_{avg}$	**TPR**$_{avg}$	**Q**
Fast-Greedy	No	56	5,36	2,19	0,76	0,09	0,11	0,91
InfoMap	No	76	3,95	2,16	0,71	0,08	0,10	0,82
Loading Eigenvector	No	61	4,92	2,16	0,73	0,08	0,10	0,86
Label Propagation	No	68	4,41	2,18	0,72	0,08	0,10	0,84
Multi-Level	No	57	5,26	2,21	0,75	0,08	0,11	0,90
Walktrap	No	74	4,06	2,16	0,77	0,07	0,09	0,86
CPM	Yes	10	4,5	3,20	0,86	0,90	1	0,21
Coda	Yes	60	4,85	2,25	0,52	0,14	0,17	0,37
Conga	Yes	49	6,20	2,18	0,84	0,08	0,10	0,92
Node-based MOGA-OCD	Yes	18	6,5	2,78	0,44	0,44	0,43	0,23
Edge-based MOGA-OCD	Yes	98	3,75	2,09	0,84	0,05	0,06	0,70

Table 5.4: Comparative assessment of community detection algorithms using the network generated from the twitter users opining about vaccines. The second column (Overlap) shows if the algorithm allows to detect overlapping communities.

Analysing these results, in the case of partitional algorithms (node overlapping is not allowed), the poor results of InfoMap algorithm can be can be clearly distinguished, achieving the lowest values to almost all metrics. The rest of algorithms obtain similar values, being them fairly low for almost of the metrics based on the network topology (CN$_{avg}$, LCC$_{avg}$ and TPR$_{avg}$)). Regarding the Q measure, the Fast-Greedy algorithm obtains the best value, and Multi-Level algorithm obtains a value that is very close. In addition, the partition detected by Fast-Greedy algorithm has a average value higher for the CN, LCC and TPR measures than the Multi-Level partitioning. Therefore, it can be concluded that the Fast-Greedy algorithm has obtained overall the best results, and this algorithm has been chosen to perform the non-overlapping community detection later on.

Continuing with the analysis of overlapping algorithms, it can be appreciated that the solutions detected by these algorithms are different from the previous ones. Edge-based MOGA-OCD algorithm identifies a much higher number of communities, but the size of these communities is also lower, grouping only three nodes in many cases. CPM algorithm detects few communities, and all of them with a high CN value. It means that this algorithm partitions the graph according to the existing triad into it. Conga algorithm achieves the highest value of Q, however for the almost the network measures this algorithm obtains poor values. Therefore, according to these measures, the algorithm which obtains better overall results is Node-based MOGA-OCD algorithm, although its modularity is low. However, this algorithm is be able to partition the graph into communities with bigger size. In addition, these communities are more structured, as shown the network metrics which all have values greater than 0.

Taking all the above into account, Fast-Greedy and Node-based MOGA-OCD algorithms have been applied using the vaccine user network to detect its communities on vaccination discussion, and to carry out and social analysis of them. Table 5.5 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 5.5 shows the most frequently re-tweet for each community, and in the first column it can be seen the results of the human-labelling process, showing if the community has a positive (P) or negative (N) opinion.

As shown this Table, 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. These results show that it is possible to identify vaccine movements from Twitter applying community detection algorithms. These algorithms are unsupervised data mining techniques, thereby human-labelling is not needed which is a big advantage for huge datasets collected from social networks such as Twitter.

The same process of human-labelling to the most frequent re-tweets and the their sentimental classification, has been carried out with the overlapping communities obtained by applying the Node-based MOGA-OCD algorithm. The results of this process are detailed in Table 5.6 where are shown 14 vaccination communities being 7 communities (1,4,5,8,9,11 and 12) opining with a positively sentiment about vaccination, against 7 which are negatively discussing (2,3,6,7,10,13 and 14). Comparing these results with the results from the non-overlapping algorithm, it can be immediately appreciated that the communities detected are smaller than the previous ones. As a consequence of this fact, the greater and most influential community identified by the first algorithm (1), in this case is divided into to communities (1 and 4). In addition the most influential user of this community is overlapped belonging to both resulting communities. The trend of anti-vaccine communities having low values related to centrality metrics is continuing. However, both types of communities (positive an negatives) have a similar have similar size for the overlapping algorithm. Regarding the overlapping users, the negative communities have a greater number of overlaps which means a higher degree of connection between them. This aspect can be very interesting to study the behaviour and relationships between the different anti-vaccine communities which through their bad comments may have a negative effect on vaccination decisions of other users. Therefore, these results show that the Node-based MOGA-OCD is able to identify vaccine opining groups and their relationships between them allowing for a better analysis of their social behaviour.

Regarding the social analysis of the communities found, each centrality metric shown in Table 5.5 provides a different aspect of its social influence. Firstly, analysing Degree Centrality, the most re-tweeted users can be identified as seen in Figure 5.4. Using this metric, users with more connections are considered as more relevant. In Figure 5.4 we can see two main communities (1 and 4) including most of the important users, or institutions, which are discussion on vaccination.

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 vaccine in the world is worth nothing if people don't use it - need social science
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 5.5: Communities detected using the 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.

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 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.

Carrying out the same social analysis with the communities discovered by the Node-based MOGA-OCD algorithm, Figure 5.5 shows the users with more connections whose are considered as more relevant for each community. As shown in this Figure, there are three positive communities (1, 4 and 12) including most of the important users discussing on vaccination to support

Id.	***Top Users***	***N.Users***	***Degree***	***Eigen.***	***Betwe.***	***Most Frequently Re-Tweet***
1(P)	***VaccinesToday***, WHO, UN, ArielPoliandri, *STcom*	7	14	1	4236,75	eu research commissioner: 'the best vaccine in the world is worth nothing if people don't use it - need social science
2(N)	**KimTaeyeonRock**, UnitedStateIX, ADavidRobinson, formulamom, USASSG_Enigma	9	2	0	106,5	#autism vaccine risks report ebook download pdf-toxic ingredients, side effects, autism, mer...
3(N)	**DHgovuk**, *KimTaeyeonRock*, davidfrum, publicroad, Rene_devries	8	2	0,01	264,85	has a medicine or vaccine made your child sick? tell us today!
4(P)	***VaccinesToday***, **BillGates**, gateshealth, sanofipasteur	4	14	1	4236,75	i couldn't agree more: vaccines are one of the cheapest ways to save lives. via @unicef
5(P)	**u38bkai**, 8r4ruinedj, 92dtenshika, tamarakeithNPR, Freebies4Mom	10	2	0	3	compound boosts effect of vaccines against hiv and flu
6(N)	**BroMuhd**, *BrotherJesse*, GiveBirthToAGod, MoneywiseMoms, Freebies4Mom	7	2	0	92,66	it was the vaccine injuries that killed my mother. - dr. horowitz
7(N)	**FactSoup**, *KimTaeyeonRock*, TheKnowIedge, Rene_devries, ggreenwald	7	2	0,01	996	a swine flu outbreak in 1976 killed one person but, the vaccine for the swine flu killed 25.
8(P)	**JodiesJumpsuit**, TabathaSouthey, CitizenRadio	3	2	0,01	1,83	there is no vaccine debate. there are just a lot of people who don't understand how vaccines work. that isn't a debate
9(P)	**allisonkilkenny**, latimes, STForeignDesk, nprnews, tamarakeithNPR	6	2	0,01	2	40% of parents are skipping or delaying their kids' vaccines for no good reason
10(N)	**DVERandy**, JoshYohe_Trib, KenTremendous, NPRHealth, *BrotherJesse*	7	2	0,01	2	miranda lambert just said v̈accines cause autism!
11(P)	**ABC**, *STcom*, STForeignDesk, seanmdav, _Dutch	6	2	0,01	99,83	cause and demand: why there is no standard ebola vaccine
12(P)	**UNICEF**, chukudebelu, Primary_Immune	3	12	0,57	3481,09	vaccines are 1 of the most effective health initiatives ever and all children should have access to them #vaccineswork
13(N)	**torieannesalt**, Disillusioned7, fsuflores	3	2	0	2	a vaccine, but for feelings
14(N)	**FeminineHygiene**, WeHateAmerica_, ActOf1871	3	6	0	3,83	good morning patriots,enjoy your #nwo supplied #chemtrails #fakedebt #fluoride...

Table 5.6: Overlapping Communities detected using the Node-based MOGA-OCD algorithm. The centrality metrics (Degree Centrality, Eigenvector Centrality and Betweenness Centrality) are related to the most influential user of the community, and they 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. Users overlaps between communities are marked in italics.

it. As in the previous case, the main users of these communities correspond to relevant health organizations. Moreover, Bill Gates who is one of the famous personalities supporting the vaccination campaigns is included in these pro-vaccine communities (4). Regarding negative vaccine communities, the Node-based MOGA-OCD algorithm detects more of them containing users less influential than using the no-overlapping algorithm. However, analysing the community overlays, it can be noticed that the negative communities have a greater number of overlapping

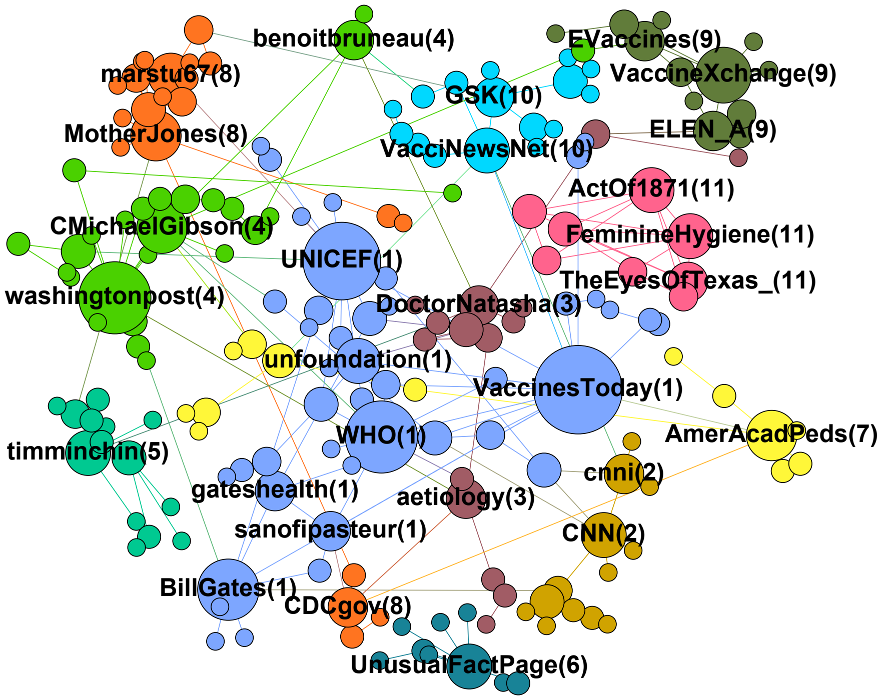


Figure 5.4: Vaccine Communities showing the **most re-tweeted** users based on their **Degree Centrality** metric (node size according to its value). Node labels are filtered by a degree value higher than 4. Top 5 users: VaccinesToday(1)(P), UNICEF(1)(P), washingtonpost(4)(P), WHO(1)(P) and BillGates(1)(P).

users showing that there is a link between them.

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 5.7. 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 5.7 shows that only one negative community (3) includes influential personalities.

In the case of overlapping communities, Figure 5.7 shows that the results related to the top 3 of most influential users is the same as in previous results. The main difference is that in these communities with overlaps, the most influential users have been separated into two smaller groups containing a user in common between them (VaccinesToday). Comparing these results with the results based on Degree Centrality, it can be appreciated that in the communities found by the Node-based MOGA-OCD algorithm the most important users for each community correspond to the most influential users. In addition any anti-vaccine group include one of the users with social influence. These findings are according to the previous analysis, where the

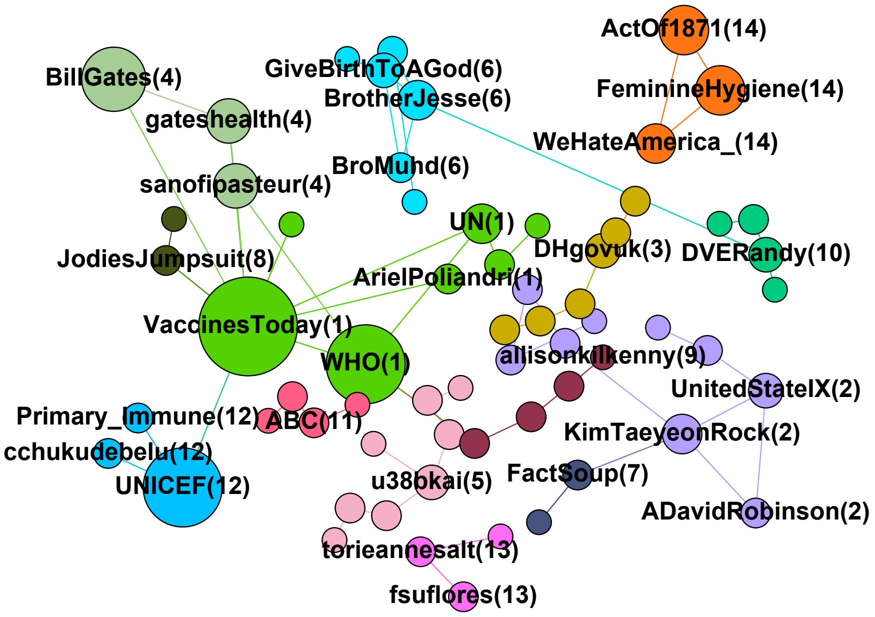


Figure 5.5: Overlapping Vaccine Communities showing the **most re-tweeted** users based on their **Degree Centrality** metric (node size according to its value). Node labels are filtered by a degree value higher than 2. Top 5 users: VaccinesToday(1)(P), UNICEF(12)(P), WHO(1)(P), BillGates(4)(P) and FeminineHygiene (14)(N).

negative communities neither have important users belonging to them.

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 5.8 shows the results obtained). Users controlling the information flow can be identified using this information. Therefore, analysing the results shown in Figure 5.8, 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.

Continuing the analysis of Betweenness Centrality, as shown in Figure 5.9, the number of users controlling the information flow is higher. Within the positive communities, BillGates appears again as a relevant user with one of the highest values in this metric. This is because he is a very popular personality all over the world, and therefore he is an important user connecting other people opinions. Communities discussing negatively on vaccination again have relevant users based on this centrality metric, as also can be seen in the results of the no-overlapping algorithm. As mentioned in different works on Social Networks [174, 153], 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.

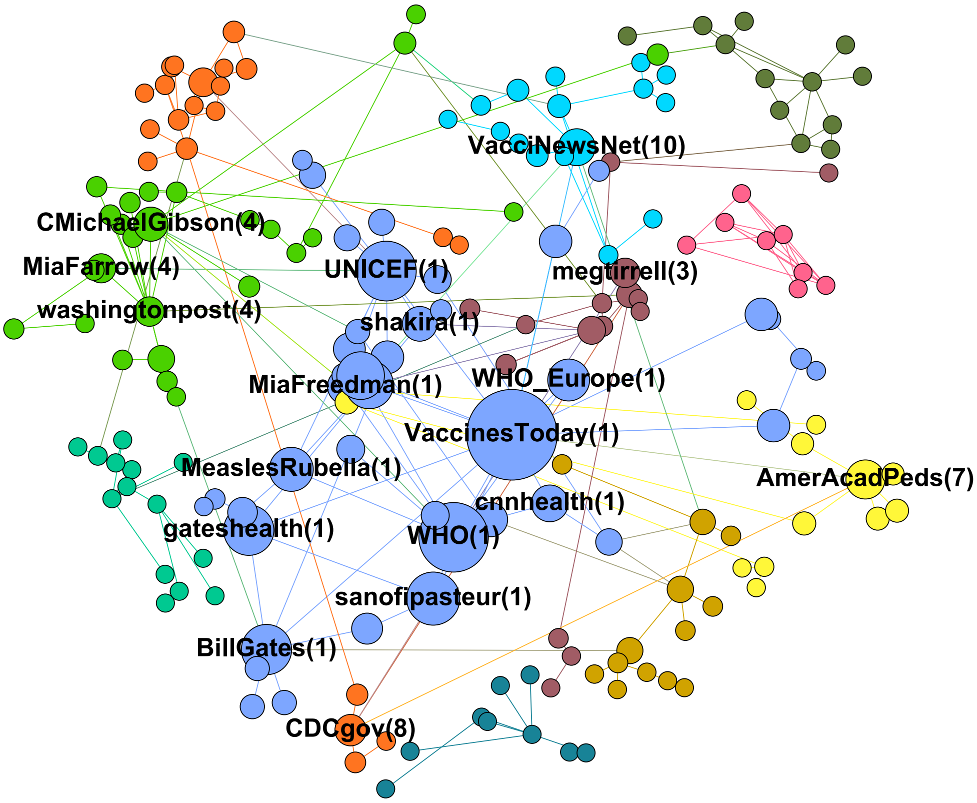


Figure 5.6: Vaccine Communities showing the **most influential** users and institutions based on the **Eigenvector Centrality** metric (node size according to its value). Node labels are filtered by a eigenvector value higher than 0,15. Top 5 users: VaccinesToday(1)(P), WHO(1), UNICEF(1)(P), BillGates(1)(P) and shakira(1)(P). Only one negative community (3) includes influential personalities.

A geographical visualization of the communities detected can help to analyse the results. Figure 5.10 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.

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. In

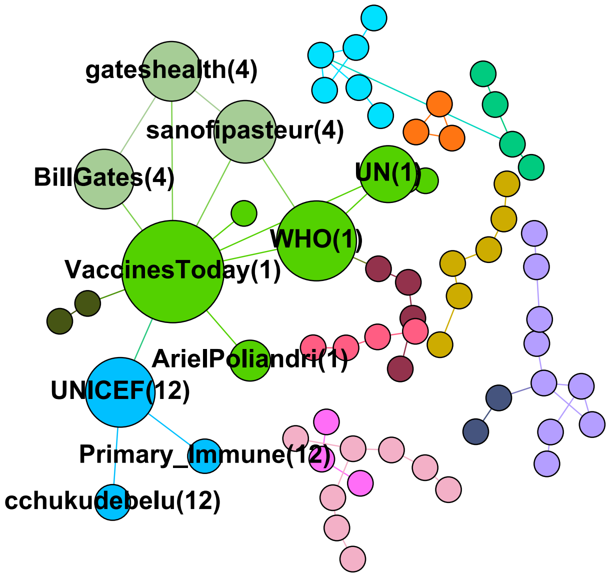


Figure 5.7: Overlapping Vaccine Communities showing the **most influential** users and institutions based on the **Eigenvector Centrality** metric (node size according to its value). Node labels are filtered by a eigenvector value higher than 0,15. Top 5 users: VaccinesToday(1)(P), WHO(1), UNICEF(13)(P), sanofipasteur(5)(P) and gateshealth(5)(P).

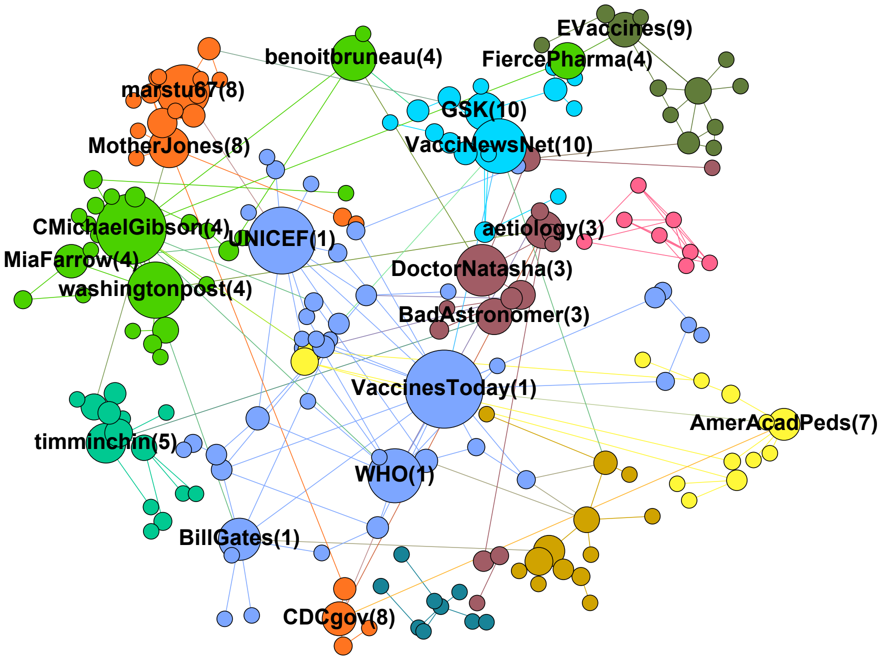


Figure 5.8: Vaccine Communities showing the users who **control the information flow** based on the **Betweenness Centrality** metric (node size according to its value). Node labels are filtered by a betweenness value higher than 1100. Top 5 users: VaccinesToday(1)(P), CMichaelGibson(4)(P), UNICEF(1)(P), washingtonpost(4)(P) and DoctorNatasha(3)(N).

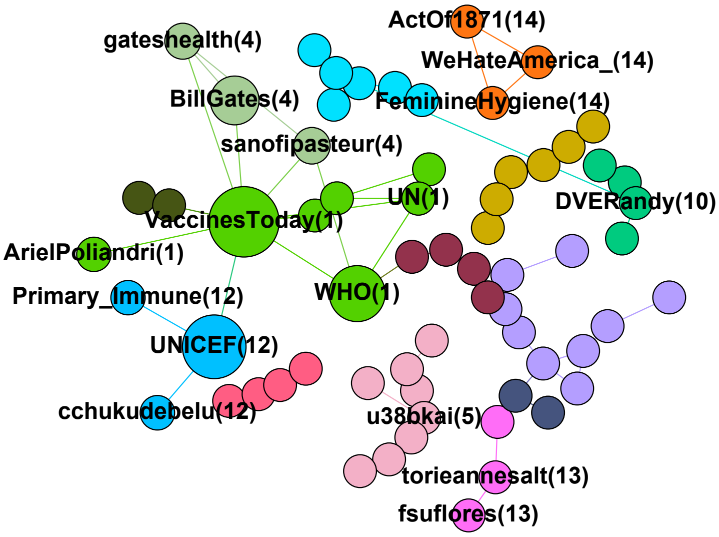


Figure 5.9: Overlapping Vaccine Communities showing the users who **control the information flow** based on the **Betweenness Centrality** metric (node size according to its value). Node labels are filtered by a betweenness value higher than 1100. Top 5 users: VaccinesToday(1)(P), UNICEF(12)(P), WHO(1)(P), BillGates (4)(P) and sanofipasteur(4)(P).

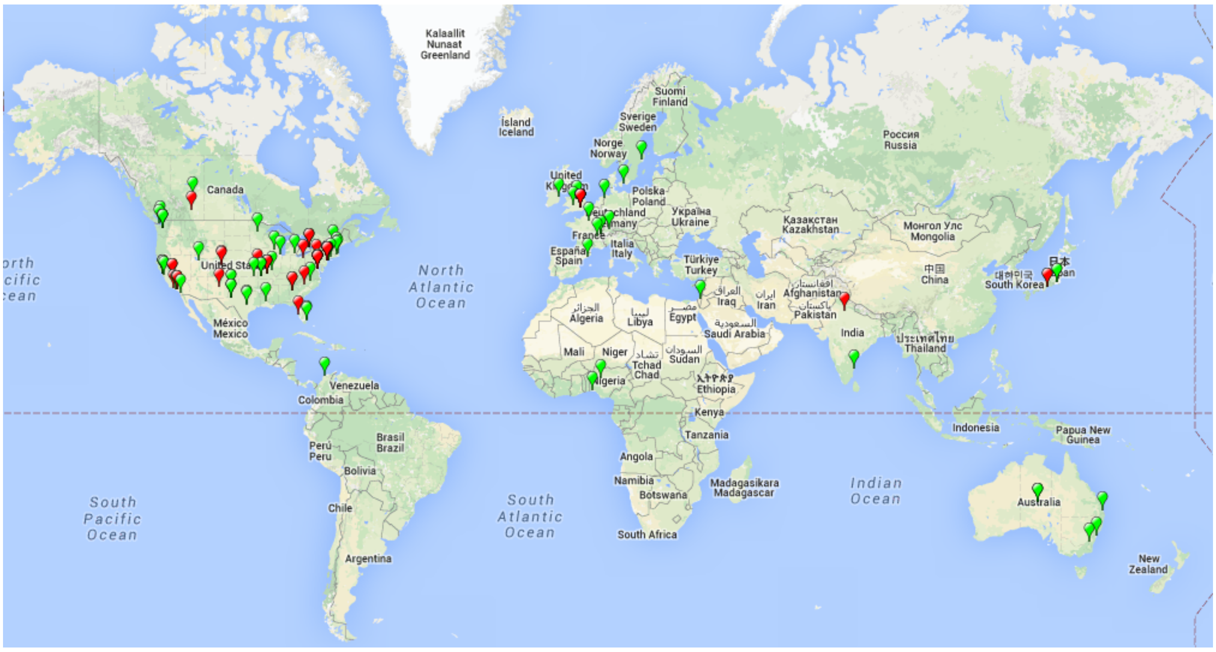


Figure 5.10: Map summarizing the location information for the communities detected using the Node-based MOGA-OCD algorithm. Anti-vaccine communities are marked in red and pro-communities in green. The most relevant anti-vaccines communities appear in EEUU.

the case of overlapping communities, the overlaps between the communities allows also the analysis of the relationships established to each other.

5.2 Conclusions

This work shows a practical application of Data Mining Techniques to detect and analyse Twitter communities which are disseminating vaccination opinions. Firstly, a preliminary analysis has been carried out showing 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 CDAs 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.

Chapter 6

Conclusions and Future Works

"Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more, so that we may fear less."

- Marie Curie

This chapter presents a discussion of the main conclusions reached in this thesis, as well as the possible future lines of work that can be carried out, in order to extend the different algorithms presented in this dissertation. Firstly, in the Conclusions section, a brief review and contributions of the different algorithms proposed in this thesis is introduced. Then all the research questions raised in the introduction section are answering according to the experimental evidences obtained by these algorithms. Finally, a summary of the possible future lines of works, based on these algorithms, is presented.

6.1 Conclusions

The main contribution of this PhD Thesis is the design of overlapping community detection algorithms, which can combine the concepts of graph clustering and genetic algorithms, in order to find new approaches to improve the classical CDAs that currently exist on the state of the art. In addition, a detailed study of several measures related to the topology structure of the networks has been carried out, analysing what are the most suitable measures to guide the search of the GAs for finding good communities. Finally, both classical algorithms and new algorithms implemented in this thesis, have been applied to a real domain, to prove how useful are the new knowledge that these algorithms can provide. In particular, the algorithm are applied to identify opinion communities related to vaccines in Twitter, analysing the influence of these communities to the rest of the users.

To achieve the main objective of this dissertation, related to the design of the new overlapping community detection algorithms combining the concepts of graph clustering with genetic algorithms, two different generations of algorithms have been implemented:

1. The first generation is focused on the design of algorithms based on a single objective Genetic Algorithm that uses a hybrid fitness function to guide the search of communities with overlaps. This generation presents three different algorithms whose main characteristics are as follows:

- **GCF-I (based on distances)**. There are two versions of this algorithm, but in both versions the fitness functions are based on distances to calculate the similarity between the graph nodes, and metrics from the graph theory:
  - **K-fixed GCF-I** uses a binary encoding where the number of communities to partitioning the graph is fixed as an input parameter.
  - **K-adaptive GCF-I** uses a more complex encoding that allows to calculate the parameter K (the number of communities of the partition) during the execution of the evolutionary process.
- **GCF-II (based on network topology metrics)**. In previous algorithms, it is necessary that each node has a set of features associated to it, which allow to measure a distance between them. However, many graphs do not have features associated to their nodes, so this new algorithm tries to identify communities optimizing metrics from the network topology of the graph.

2. The second generation is based on Multi-Objective Genetic Algorithms, where the fitness function implemented optimizes two different objective functions; the first one that is used to maximize the internal connectivity of the communities, and the second one that is used to minimize the external connections to the rest of the graph. In this new generation, two different algorithms have been designed and implemented, and their main differences are related to the encoding used to represent the individuals:

   - **Node-based MOGA-OCD** where the alleles of the individuals represent the nodes of the graph.
   - **Edge-based MOGA-OCD** where a new encoding scheme based on the edges is used.

   In both algorithms, the value of K (number of communities in the graph) has been directly encoded as part of the chromosome.

Taking into account all the experimental analysis carried out for each algorithms proposed in this thesis, it is possible to provide an answer to the main research questions that have been briefly presented in the Introduction section. Next, a revision of these research questions is carried out, and the answers are discussed on the basis of the experimental conclusions reported in this dissertation:

- **Q1:** *Is it possible to combine clustering methods based on distances with genetic graph-based approaches to improve the results obtained by classical overlapping community detection algorithms?*

  The classical clustering algorithms, such as K-means, typically use distance measures to measure the similarity between the elements of the dataset for grouping them. On the other hand, the classical CDAs are typically based on the topology information of the graph to partition it.

  Therefore, trying to validate if these two different methodologies can be combined to improve the detection of overlapping communities within a graph, the two first algorithms designed in the first generation have been implemented (K-fixed GCF-I and K-adaptive

GCF-I). These algorithms are based on classical GAs with a single optimization criteria, and in order to combine both methodologies, different fitness functions have been designed combining distance metrics between the nodes, with metrics from the graph theory to guide the search. The fitness functions based on distances will try to optimize the the quality of the communities detected by minimizing the distance between the nodes which belong to a same cluster, whereas the distance between the centroids of the communities will be maximized.

The experimental findings obtained for these algorithms (see section 3.1.3), show that the hybrid fitness function combining the distance between nodes (intra and inter cluster distance), and the CC, is able to reach better results than the other classical CDAs studied. In addition, from an overall analysis of all the studied metrics, it can be noticed that the new K-adaptive GCF-I algorithm improves the results in all the cases comparing to the results obtained by the K-fixed GCF-I algorithm. This new evolutionary approach finds communities that have an appropriate size, reduced overlapping and closer distances between the nodes.

Comparing K-adaptive GCF algorithm against CPM and EBC in more detail, it can be noticed that the communities detected by the genetic algorithms are smaller than the communities identified by the classical CDAs. In addition, it is important to observe that each community generated by the genetic algorithm is contained, or partially contained, in a community generated by the CPM algorithm. It means that the genetic algorithm has tuned up the original community definition of this classical algorithm.

- **Q2:** *Are these algorithms able to identify quality communities only using network topology measures as objective functions?*

The third algorithm from the first generation has been inspired by network topology analysis (GCF-II), and it is based on the use of network measures (Density, Centralization, Heterogeneity, Neighbourhood, Clustering Coefficient) to guide the search. In order to choose the better measures for the fitness function, a comparative assessment of network measures has been carried out, and then this algorithm is compared to other CDAs, and the previous versions of the algorithm (GCF-I).

The analysis of network metrics shown that Density obtains the better results, but the communities detected are small and no-overlapping. On the other hand, using the Heterogeneity measure as fitness function, the algorithm is able to identify communities with a good size and appropriate overlapping. Therefore, both measures have been combined in the weighted fitness function of the algorithm to detect communities with good quality considering all the features.

Finally, the results obtained by K-adapted GCF-I and GCF-II algorithms have been compared against the results of CPM algorithm. The experimental results show that the new version of the algorithm (GCF-II) improves the results in all of the cases studied. It means that using a fitness function based only on measures related to the network topology, the algorithm detects communities with an appropriate size, reduced overlapping, members very connected, and close distances between the different communities detected. However, the results show how the changes in the network structure of the dataset used, affects to the quality of solutions detected by the all the algorithms. Therefore, it can be concluded that these algorithms are dependent on the network topology.

- **Q3:** *How can multi-objective genetic algorithms deal with the problem of dependence on the graph structure shown by the classic algorithms?*

  In order to deal with the problem of the dependence on the network topology detected previously in the experimental phase of the first generation of the algorithms (based on a genetic algorithms with a single objective), a second generation of algorithms have been designed using a multi-objective approach.

  In CDP, it can be difficult to define what is a good community, because it can depends on the application domain. However, according to the own structure of the graphs, a good community can be defined as well connected internally, whereas it should be well separated from the rest of the nodes of the graph. Taking into account these two features, the Multi-Objective Genetic Algorithms will allows to find solutions optimizing simultaneously both objectives using a Pareto-Optimality Frontier.

  Currently, there are several studies in the literature showing that the use of the edge information to detect communities into graphs is a good methodology. Particularly, when these methods are applied to real world networks which tend to be sparse, and the node-based methods often have difficulties finding large size communities. For this reason, in the second generation of the algorithms of these thesis, two different algorithms (MOGA-OCD) have been implemented whose main difference is based on the encoding. The first algorithm (Node-based MOGA-OCE) uses an encoding where the individual represents groups of nodes, whereas in the second algorithm the individual represents edges of the graph.

  Both algorithms have been evaluated against several overlapping algorithms from the state of the art, using two different types of datasets: real world networks providing the ground truth of their communities to evaluate the accuracy; and more complex real networks without ground truth to evaluate the effectiveness and quality of detected solutions.

  Analysing the experimental results obtained with all the datasets, some overall conclusions can be drawn. The classical CDAs usually have good results when the graph is very structured, and the algorithm can partition it according to its network topology. But when these algorithms are applied using large size graphs, they have problems to partitioning it into quality communities. In addition, for unstructured or sparse graphs, the accuracy and quality of the communities detected by these algorithms significantly decrease. On the other hand, using unstructured or sparse graphs, the node-based MOGA-OCD algorithm improves the results in almost all the cases. Related to structured graphs, this version of the algorithm obtains similar results, or closer, to the rest. However, the edge-based MOGA-OCD algorithm obtains good results for both types of graphs (structured and unstructured).

  So,in general terms, the MOGA-OCD algorithms obtain good results for the different types of graphs, including largest datasets where in some cases the classic algorithms are not able to compute even a graph partition. Furthermore, using the dataset collection with ground-truth, these proposed algorithms reaches the best accuracy in almost all cases.

- **Q4:** *How can these algorithms be applied to real social domains to acquire useful information?*

  The last part of the dissertation presents a practical CDAs application to discover and analyse Twitter communities, which are disseminating vaccination opinions. For this purpose a dataset collected from Twitter, and the vaccination coverage rates retrieved from the immunization monitoring system of WHO, have been used to perform several analysis.

Using both datasets, an initial study has been carried out focused on measuring the potential influence of vaccine opinions based on the variation in the coverage rates. 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.

Afterwards, a network representation of the Twitter dataset based on the user re-tweets is generated to apply over it the CDAs extracting the user communities talking about vaccines. The detailed analysis of these communities shows that the most important, and influential users belonging to these communities, are 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.

Therefore, taking into account all the experimental results presented, it can be concluded that the application of CDAs is useful, and they can be used to find and track vaccine movements, discovering new knowledge in data that could be useful to improve Public Healthcare strategies about immunization. Moreover, this new knowledge can also be used to detect and locate communities against vaccination, which could generate future disease outbreaks in different parts of the world.

6.2 Future Works

Finally, there are several lines of work that could be extended in the near future related to the different algorithms and applications presented in this dissertation:

- It could be interesting to extend the algorithms designed in this thesis, in order to handle very large size networks, using Big Data frameworks for massive data processing such as Apache Hadoop or Spark.
- Related to the connectivity measures used as objective functions in the fitness function, it could be important to carry out a more detailed analysis of all of them to study which one is the more suitable for each different graph topology.
- It could be interesting to develop a cooperative co-evolutionary approach of the algorithms, dividing the problem into two subcomponents: one sub-population to search for an optimal exclusive partition of the graph, and other to search for the optimal overlapping between these communities.
- Due to the number of different connectivity measures that could used to guide the search of solutions, it could be very interesting to design a self-adaptive version of the algorithms for selecting the optimization criteria used as fitness function during the evolving process.

Chapter 7

Conclusiones y Trabajos Futuros

"En la vida, no hay nada que temer, solo hay que comprender.
Ahora es el momento de entender más, para que podamos temer menos."

- Marie Curie

Este capítulo presenta las principales conclusiones alcanzadas en esta tesis, así como las posibles líneas de trabajo futuras que se pueden llevar a cabo, con el fin de ampliar los diferentes algoritmos presentados en la tesis. En primer lugar, en la sección de conclusiones, se realiza un breve resumen de los diferentes algoritmos propuestos en esta tesis, así como de sus contribuciones. Seguidamente, se pasa a responder las preguntas de investigación planteadas en la sección de introducción, a través de las evidencias experimentales obtenidas por los diferentes algoritmos implementados. Finalmente, se plantean cuales podrías ser las posibles futuras líneas de trabajo, que se podrían realizar sobre estos algoritmos para mejorarlos.

7.1 Conclusiones

La principal contribución de esta tesis doctoral es el diseño e implementación de nuevos algoritmos de detección de comunidades solapadas, combinando los conceptos de las técnicas de clustering de grafos y algoritmos genéticos. El objetivo principal de estos nuevos algoritmos será encontrar nuevos enfoques para mejorar los algoritmos clásicos que existen actualmente en el estado del arte. Además, se ha realizado un estudio detallado de varias medidas relacionadas con la estructura topológica de las redes, analizando cuáles son las más adecuadas para guiar los algoritmos genéticos propuestos, y así realizar buenas divisiones de los grafos en comunidades. Finalmente, tanto los algoritmos clásicos como los nuevos algoritmos implementados en esta tesis, se han aplicado a un dominio real, para demostrar la utilidad del nuevo conocimiento que estos algoritmos pueden proporcionar. En particular, dichos algoritmos se han aplicado para identificar comunidades de opinión relacionadas con las vacunas en Twitter, analizando la posible influencia de ellas sobre el resto de los usuarios, y estudiando un posible efecto sobre las tasas de vacunación.

Para lograr el objetivo principal de esta tesis se han implementado dos generaciones diferentes de algoritmos:

1. La primera generación se centra en el diseño de algoritmos basados en un algoritmo genético

estándar con un solo objetivo, y que utiliza una función de aptitud híbrida para guiar la búsqueda de las comunidades con solapamientos. Esta generación presenta tres algoritmos diferentes cuyas principales características son las siguientes:

- **GCF-I (basado en distancias)**. Hay dos versiones de este algoritmo, pero en ambas versiones las funciones de aptitud se basan en distancias para calcular la similitud entre los nodos del grafo, así como en métricas relacionadas con la teoría de los grafos:
  - El algoritmo **K-fixed GCF-I** utiliza una codificación binaria donde el número de comunidades en las que dividir el grafo se fija como un parámetro de entrada.
  - El algoritmo **K-adaptive GCF-I** utiliza una codificación más compleja que permite calcular el parámetro K (úmero de comunidades de la partición) durante la ejecución del propio proceso evolutivo de forma automática.
- **GCF-II (basado en métricas de topología de la red)**. En algoritmos anteriores, es necesario que cada nodo tenga un conjunto de características asociadas a él, lo que va permite medir distancias o similitudes entre ellos. Sin embargo, muchos grafos reales no tienen características asociadas a sus nodos, por lo que este nuevo algoritmo intenta identificar comunidades solamente utilizando métricas relacionadas con la propia topología de la red o grafo.

2. La segunda generación de algoritmos implementada se basa en Algoritmos Genéticos Multi-Objetivo, donde la función de aptitud implementada optimiza dos funciones objetivos diferentes simultáneamente; tratando el primer objetivo de maximizar la conectividad interna de las comunidades, y el segundo tratando de minimizar las conexiones externas con el resto de grafo. En esta nueva generación se han diseñado e implementado dos algoritmos diferentes, cuyas principales diferencias están relacionadas con la codificación utilizada para representar a los individuos:

   - **Node-based MOGA-OCD** donde los alelos de los individuos representan los nodos del grafo.
   - **Edge-based MOGA-OCD** donde se utiliza un nuevo esquema de codificación basado en los enlaces del grafo.

   En ambos algoritmos, el valor de K (número de comunidades a detectar) se ha codificado directamente como parte del cromosoma.

Teniendo en cuenta todos los análisis experimentales realizados para cada uno de los algoritmos propuestos en esta tesis, es posible dar una respuesta a las principales preguntas de investigación presentadas en la sección de Introducción. A continuación se realiza una revisión de estas preguntas de investigación, argumentando sus respuestas en base a las conclusiones experimentales presentadas en los distintos capítulos de la tesis:

- **Q1:** *¿Es posible combinar métodos de agrupación (clustering) basados en distancias, con enfoques genéticos basados en grafos para mejorar los resultados obtenidos por los algoritmos clásicos de detección de comunidades solapadas?*

  Los algoritmos clásicos de agrupamiento (clustering), como K-means, suelen utilizar métricas de distancias para medir la similitud entre los elementos de un conjunto de datos dado, y mediante esta información formar grupos de elementos similares. Por otro lado, los algoritmos de detección de comunidades clásicos (CDAs), utilizan típicamente la propia información de la topología del grafo para dividirlo.

Por lo tanto, para intentar validar si estas dos metodologías diferentes pueden combinarse para mejorar los métodos de detección de comunidades superpuestas, se han implementado los dos primeros algoritmos pertenecientes a la primera generación de esta tesis (K-fixed GCF-I y K-adaptive GCF-I). Estos algoritmos están basados en Algorithmos Genéticos (GAs) clásicos con un solo criterio de optimización, y para combinar ambas metodologías, se han diseñado diferentes funciones de aptitud (fitness) combinando métricas de distancia entre los nodos, con métricas de la teoría de grafos para guiar la búsqueda. Las funciones de aptitud basadas en distancias intentarán optimizar la calidad de las comunidades detectadas minimizando la distancia entre los nodos que pertenecen a un mismo grupo, mientras que se maximizará la distancia entre los centros de las distintas comunidades.

Los resultados experimentales obtenidos para ambos algoritmos (ver sección 3.1.3) muestran que la función de aptitud híbrida combinando las distancias (entre nodos y centros) y el coeficiente de clusterización (CC), es capaz de alcanzar mejores resultados que los otros CDAs clásicos estudiados. Además, a partir de un análisis global de todas las métricas estudiadas, se puede observar que el nuevo algoritmo K-adaptive GCF-I mejora los resultados en todos los casos en comparación con los resultados obtenidos por el algoritmo K-fijo GCF-I. Este nuevo enfoque evolutivo encuentra comunidades que tienen un tamaño adecuado, reduciendo la superposición y distancias más cercanas entre los nodos.

Al comparar el algoritmo K-adaptive GCF-Icon CPM y EBC más detalladamente, se puede observar que las comunidades detectadas por los algoritmos genéticos son más pequeñas que las comunidades identificadas por los CDA clásicos. Además, es importante observar que cada comunidad generada por el algoritmo genético está contenida, o parcialmente, en una comunidad generada por el algoritmo CPM. Esto significa que el algoritmo genético ha afinado la definición original de la comunidad de este algoritmo clásico, detectando comunidades más compactas.

- **Q2:** *¿Son estos algoritmos capaces de identificar comunidades de calidad sólo usando medidas sobre la topología de la red como funciones objetivo?*

El tercer algoritmo implementado en la primera generación se ha inspirado en el análisis de la topología del grafo (GCF-II), y se basa en el uso de medidas de conectividad de la red (Densidad, Centralización, Heterogeneidad, y Coeficiente de Clusterización) para guiar la búsqueda. Con el fin de elegir las métricas que se adapten mejor a resolver el problema para fijarlas como criterios de optimización de la función de aptitud del algoritmo, se realiza una evaluación comparativa de todas las medidas de la red mencionadas. Finalmente el nuevo algoritmo es comparado con otros CDAs, y las versiones anteriores del mismo (GCF-I).

El análisis de las métricas de red muestra que la Densidad obtiene mejores resultados en cuanto a distancias, pero las comunidades detectadas son pequeñas y no suelen tener solapes. Por otro lado, utilizando la medida de Heterogeneidad, el algoritmo es capaz de identificar comunidades con un buen tamaño y la superposición adecuada. Por lo tanto, ambas medidas se han combinado en la función de aptitud ponderada del algoritmo para detectar comunidades con buena calidad considerando todas las características.

Por último, los resultados obtenidos por K-adaptive GCF-I y GCF-II son comparados con los resultados del algoritmo CPM. Los resultados experimentales muestran que la nueva versión del algoritmo (GCF-II) mejora los resultados en todos los casos estudiados. Esto significa que utilizando una función de aptitud que se base solamente en medidas relacionadas con la topología de la red, el algoritmo detecta comunidades con un tamaño apropiado, solapamiento reducido, miembros muy conectados y distancias cercanas entre las diferentes comunidades detectadas. Sin embargo, los resultados muestran cómo los cambios en la estructura de red del conjunto de datos utilizados, afecta a la calidad de las soluciones detectadas por todos los algoritmos. Por lo tanto, se puede concluir que estos algoritmos son dependientes de la topología de red.

- **Q3:** *¿Cómo pueden los algoritmos genéticos multi-objetivo lidiar con el problema de la dependencia de la estructura de los grafos mostrada por los algoritmos clásicos?*

Con el fin de abordar el problema de la dependencia de la topología de red detectada previamente en la fase experimental de la primera generación de los algoritmos (basada en algoritmos genéticos con un solo objetivo), se ha diseñado una segunda generación de algoritmos utilizando un enfoque multi-objetivo.

En el problema de detección de comunidades, puede ser difícil definir lo que es una buena comunidad, ya que puede depender del dominio de la aplicación. Sin embargo, según la propia estructura de los grafos, de una forma general se podría decir que una buena comunidad puede ser aquella que esté bien conectada internamente, y poco conectada con el resto de los nodos del grafo (separada del resto de comunidades). Teniendo en cuenta estas dos características, los Algoritmos Genéticos Multi-Objetivos permitirán encontrar soluciones que optimicen simultáneamente ambos objetivos utilizando una Frontera de Pareto-Optimalidad.

Actualmente, existen varios estudios en el estado del arte que muestran que el uso de la información sobre los enlaces para detectar comunidades en un grafo es una buena metodología. Particularmente, cuando estos métodos se aplican a redes que representan dominios del mundo real que tienden a ser dispersas, y los métodos basados en nodos a menudo tienen dificultades para encontrar comunidades de gran tamaño. Por esta razón, en la segunda generación de los algoritmos de esta tesis, se han implementado dos algoritmos diferentes (MOGA-OCD) cuya principal diferencia se basa en la codificación. El primer algoritmo (Node-based MOGA-OCD) utiliza una codificación donde el individuo representa grupos de nodos, mientras que en el segundo algoritmo el individuo representa los enlaces del grafo.

Ambos algoritmos han sido evaluados contra varios CDAs clásicos que permiten solapamiento, utilizando dos tipos diferentes de conjuntos de datos: redes reales que proporcionan un etiquetado de sus comunidades para evaluar la precisión; Y redes reales más complejas sin etiquetado para evaluar la efectividad y calidad de las soluciones detectadas.

Al analizar los resultados experimentales obtenidos con todos los conjuntos de datos, se pueden extraer algunas conclusiones generales. Los CDA clásicos suelen tener buenos resultados cuando el grafo está muy estructurado, y el algoritmo puede dividirlo según su topología de red. Pero cuando estos algoritmos se aplican utilizando grafos de gran tamaño, tienen problemas para dividirlo en comunidades de calidad. Además, para grafos no estructurados o dispersos, la precisión y calidad de las comunidades detectadas por estos algoritmos disminuyen significativamente. Por otro lado, utilizando algoritmos no estructurados o dispersos, el nuevo algoritmo MOGA-OCD basado en nodos mejora los resultados en casi todos los casos. En relación con los grafos estructurados, esta versión del algoritmo obtiene resultados similares, o muy cercanos al resto. Sin embargo, el algoritmo basado en enlaces obtiene buenos resultados para ambos tipos de grafos (estructurado y no estructurado).

Por lo tanto, en términos generales, se puede concluir que los algoritmos MOGA-OCD obtienen buenos resultados para los diferentes tipos de grafos, incluyendo conjuntos de datos de mayor tamaño donde en algunos casos los algoritmos clásicos no son capaces de realizar una partición del mismo. Además, utilizando la colección de conjuntos de datos etiquetados (se indica la información de a qué comunidad debe pertenecer cada nodo), los nuevos algoritmos propuestos alcanzan una mayor precisión en casi todos los casos estudiados.

- **Q4:** *¿Cómo se pueden aplicar estos algoritmos a dominios sociales reales para adquirir información útil?*

La última parte de la tesis presenta una aplicación práctica de los algoritmos de detección de comunidades para descubrir y analizar comunidades en Twitter, que están difundiendo opiniones sobre la vacunación. Para ello se ha utilizado un conjunto de datos recopilado de Twitter, así como información sobre las tasas de vacunación extraídas de la web de la Organización Internacional de la Salud (WHO).

Usando ambos conjuntos de datos, se ha realizado un estudio inicial enfocado a medir la influencia potencial que pueden estar teniendo las opiniones sobre las vacunas diseminadas en Twitter, en las tasas de vacunación de los países. Los resultados obtenidos en este análisis preliminar muestran, que en algunos casos sí que podría estar afectando las opiniones de vacunas de los usuarios de Twitter, en el proceso de toma de decisiones de vacunación de algunos países. Sin embargo, también se puede apreciar que la mayoría de las comunidades detectadas no están en contra de las vacunas. De hecho, actualmente la mayoría de los movimientos emergidos en Twitter sobre este tema están intentando apoyar la vacunación, tratando de que haya un incremento de las tasas para tener una cobertura lo mayor posible.

Posteriormente, se genera una representación en forma de grafo con los datos extraídos de Twitter basado en los re-tweets que se hacen entre los distintos usuarios. Sobre esta grafo se van a aplicar distintos algoritmos de detección de comunidades, para identificar aquellas comunidades están hablando sobre las vacunas y tienen opiniones similares. El análisis detallado de estas comunidades muestra que los usuarios más importantes e influyentes incluidos en ellas, están apoyando el movimiento de vacunación. Mientras que las comunidades de vacunas negativas a menudo incluyen pocos usuarios, y que además no están bien conectados. Finalmente, una visualización geográfica de estas comunidades muestra que los países más relevantes (Irlanda, Reino Unido, Canadá y Australia) hablando del tema están apoyando la vacunación. Por otro lado, la mayoría de las comunidades que difunden opiniones negativas se encuentran en EEUU.

Por lo tanto, teniendo en cuenta todos los resultados experimentales presentados, se puede concluir que la aplicación de CDAs es útil, y pueden ser utilizados para encontrar y rastrear los movimientos de vacunas, aportando nuevo conocimiento que podrían ser útil para mejorar las estrategias de Salud Pública sobre la inmunización. Además, este nuevo conocimiento también puede utilizarse para intentar detectar y localizar donde podrían generarse futuros brotes de enfermedades en diferentes partes del mundo, debido a que las tasas de vacunación vayan a disminuir.

7.2 Trabajos Futuros

Por último, existen varias líneas de trabajo que podrían ampliarse en un futuro próximo en relación con los diferentes algoritmos y aplicaciones presentados en esta tesis:

- Podría ser interesante extender los algoritmos diseñados en esta tesis, con el fin de manejar grafos de gran tamaño, utilizando herramientas de Big Data para el procesamiento masivo de datos como Apache Hadoop o Spark.
- En relación con las medidas de conectividad utilizadas como criterios de optimización en la función de aptitud, podría ser importante realizar un análisis más detallado de todas ellas para estudiar cuál es la más adecuada para cada topología de grafo diferente.
- Podría ser interesante desarrollar un enfoque co-evolutivo cooperativo de los algoritmos, dividiendo el problema en dos subcomponentes: una subpoblación para buscar una partición óptima del gráfico, y otra para buscar la superposición óptima entre estas comunidades.
- Debido a que existen varias medidas diferentes de conectividad que podrían utilizarse para guiar la búsqueda de soluciones, podría ser muy interesante diseñar una versión autoadaptable de los algoritmos. De esta manera, durante el propio proceso evolutivo, se seleccionaría automáticamente los criterios de optimización que sean mejores para encontrar mejores soluciones.

Bibliography

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

[2] Eurovision song contest, 2011. http://www.eurovision.tv.

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

[4] Twitter web site, 2013. twitter.com.

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

[6] H. Adeli and K. Sarma. *Cost Optimization of Structures: Fuzzy Logic, Genetic Algorithms, and Parallel Computing.* John Wiley & Sons, 2006.

[7] J. Aguilar. Resolution of the clustering problem using genetic algorithms. *International Journal of Computers*, 1(4):237 – 244, 2007.

[8] Y. Ahn, J. Bagrow, and S. Lehmann. Communities and hierarchical organization of links in complex networks, 2010. *Nature*, 466:761.

[9] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

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

[11] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576. Association for Computational Linguistics, 2011.

[12] A. Arenas. Interactions in pretty good privacy, 2015 (accessed May 3, 2016).

[13] S. Asur and 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*, volume 1, pages 492–499. IEEE, 2010.

[14] T. Back, D. B. Fogel, and Z. Michalewicz, editors. *Handbook of Evolutionary Computation.* IOP Publishing Ltd., Bristol, UK, UK, 1997.

[15] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, March 2004.

[16] I. Batal, H. Valizadegan, G. F. Cooper, and M. Hauskrecht. A temporal pattern mining approach for classifying electronic health record data. *ACM Trans. Intell. Syst. Technol.*, 4(4):63:1–63:22, Oct. 2013.

[17] G. Bello, H. Menéndez, S. Okazaki, and D. Camacho. Extracting collective trends from twitter using social-based data mining. In *Computational Collective Intelligence. Technologies and Applications*, pages 622–630. Springer, 2013.

[18] G. Bello-Orgaz and D. Camacho. Evolutionary clustering algorithm for community detection using graph-based information. In *Evolutionary Computation (CEC), 2014 IEEE Congress on*, pages 930–937. IEEE, 2014.

[19] G. Bello-Orgaz, J. Hernandez-Castro, and D. Camacho. A survey of social web mining applications for disease outbreak detection. In *Intelligent Distributed Computing VIII*, pages 345–356. Springer International Publishing, 2015.

[20] G. Bello-Orgaz, J. Hernandez-Castro, and D. Camacho. Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems*, 66:125–136, 2016.

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

[22] G. Bello-Orgaz, H. Menendez, and D. Camacho. Adaptive k-means algorithm for overlapped graph clustering. *International Journal of Neural Systems*, 22(05):1250018 1–19, 2012.

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

[24] H.-G. Beyer and H.-P. Schwefel. Evolution strategies - a comprehensive introduction. 1(1):3–52, 2002.

[25] J. C. Bezdek, J. Keller, R. Krisnapuram, and N. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing (The Handbooks of Fuzzy Sets)*. Springer, 1 edition, Mar. 2005.

[26] X. Bin, W. Min, L. Yanming, and F. Yu. Improved genetic algorithm research for route optimization of logistic distribution. In *Proceedings of the 2010 International Conference on Computational and Information Sciences*, ICCIS '10, pages 1087–1090, Washington, DC, USA, 2010. IEEE Computer Society.

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

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

[29] N. Boccara. *Modeling Complex Systems*. Springer, 1 edition, 2003.

[30] T. Bodnar and M. Salathé. Validating models for disease detection using twitter. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 699–702. International World Wide Web Conferences Steering Committee, 2013.

[31] I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo. The maximum clique problem. In *Handbook of Combinatorial Optimization*, pages 1–74. Kluwer Academic Publishers, 1999.

[32] T. Botsis, M. D. Nguyen, E. J. Woo, M. Markatou, and 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):631–638, 2011.

[33] S. Brien, N. Naderi, A. Shaban-Nejad, L. Mondor, D. Kroemker, and 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*, pages 683–686. International World Wide Web Conferences Steering Committee, 2013.

[34] C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, Sept. 1973.

[35] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl. Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS medicine*, 5(7):e151, 2008.

[36] H. A. Carneiro and E. Mylonakis. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*, 49(10):1557–1564, 2009.

[37] H. Chen and D. Zeng. Ai for global disease surveillance. *Intelligent Systems, IEEE*, 24(6):66–82, 2009.

[38] W. Chen, C. Wang, and 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*, pages 1029–1038. ACM, 2010.

[39] Y.-Y. Chen and K.-Y. Young. An som-based algorithm for optimization with dynamicc weight updating. *International Journal of Neural Systems*, 17(3):171 – 181, 2007.

[40] A. Clauset. Finding local community structure in networks. *Phys. Rev. E*, 72:026132, Aug. 2005.

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

[42] C. A. C. Coello, D. A. Van Veldhuizen, and G. B. Lamont. *Evolutionary algorithms for solving multi-objective problems*, volume 242. Springer, 2002.

[43] A. M. Cohen and W. R. Hersh. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71, 2005.

[44] Coley. *An Introduction to Genetic Algorithms for scientists and engineers.* World Scientific Publishing, 1999.

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

[46] N. Collier, S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q.-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, et al. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–2941, 2008.

[47] N. Collier, R. M. Goodwin, J. McCrae, S. Doan, A. Kawazoe, M. Conway, A. Kawtrakul, K. Takeuchi, and D. Dien. An ontology-driven system for detecting global health events. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 215–222. Association for Computational Linguistics, 2010.

[48] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM, 2010.

[49] L. Danon, A. Diaz-Guilera, and A. Arenas. The effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(11):P11010, 2006.

[50] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

[51] S. Das, A. Abraham, and A. Konar. Automatic clustering using an improved differential evolution algorithm. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 38(1):218 –237, jan. 2008.

[52] A. Davis. Southern club women, 2011 (accessed May 3, 2016).

[53] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *Evolutionary Computation*, 6(2):182–197, 2002.

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

[55] M. Dehmer, editor. *Structural Analysis of Complex Networks.* Birkhäuser Publishing, 2010. in press.

[56] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[57] J. Dong and S. Horvath. Understanding network concepts in modules. *BMC systems biology*, 1(1):1, 2007.

[58] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW.* Oxford University Press, 2003.

[59] EBU. http://www.ebu.ch/. October 2010.

[60] A. E. Eiben and J. E. Smith. *Introduction to Evolutionary Computing.* Natural Computing. Springer-Verlag Berlin Heidelberg, 2003.

[61] S. S. Elisa. Survey: Graph clustering. *Comput. Sci. Rev.*, 1(1):27–64, Aug. 2007.

[62] D. Eppstein, M. Löffler, and D. Strash. Listing all maximal cliques in large sparse real-world graphs. *J. Exp. Algorithmics*, 18:3.1:3.1–3.1:3.21, Nov. 2013.

[63] L. J. Eshelman, R. A. Caruana, and J. D. Schaffer. Biases in the crossover landscape. In *Proceedings of the Third International Conference on Genetic Algorithms*, pages 10–19, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[64] T. Evans and R. Lambiotte. Line graphs, link partitions, and overlapping communities. *Physical Review E*, 80(1):016105, 2009.

[65] L. Fan and H. Meng. Application research on optimal path based on genetic algorithm. *JDCTA*, 4(8):199–202, 2010.

[66] D. Fenn, O. Suleman, J. Efstathiou, and N. Johnson. How does europe make its mind up? connections, cliques, and compatibility between countries in the eurovision song contest. *Physica A: Statistical Mechanics and its Applications*, 360(2):576–598, February 2005.

[67] D. H. Fischer's. Graph representing colonial american dissidents, 2015 (accessed May 3, 2016).

[68] 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*, volume 130, pages 16–28. The Royal Society, 1930.

[69] M. Fisichella, A. Stewart, A. Cuzzocrea, and K. Denecke. Detecting health events on the social web to enable epidemic intelligence. In *String Processing and Information Retrieval*, pages 87–103. Springer, 2011.

[70] D. Fogel and L. Fogel. An introduction to evolutionary programming. In *Artificial Evolution*, volume 1063 of *Lecture Notes in Computer Science*, pages 21–33. Springer Berlin Heidelberg, 1996.

[71] L. J. Fogel. Autonomous automata. *Industrial Research*, 4:14–19, 1962.

[72] 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):243, 1999.

[73] S. Fortunato, V. Latora, and M. Marchiori. Method to find community structures based on information centrality. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 70(5):056104, 2004.

[74] L. C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[75] A. A. Freitas. A review of evolutionary algorithms for data mining. In *In: Soft Computing for Knowledge Discovery and Data Mining*, pages 61–93, 2007.

[76] D. Gatherer. Birth of a meme: The origin and evolution of collusive voting patterns in the eurovision song contest. *Journal of Memetics-Evolutionary Models of Information Transmission*, 8(1):28–36, 2004.

[77] D. Gatherer. Comparison of eurovision song contest simulation with actual results reveals shifting patterns of collusive voting alliances. *Journal of Artificial Societies and Social Simulation*, 9(2):1, 2006.

[78] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[79] V. Ginsburgh and A. Noury. The eurovision song contest. is voting political or cultural? *European Journal of Political Economy*, 24(1):41–52, 2008.

[80] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[81] F. Godlee, J. Smith, and H. Marcovitch. Wakefield's article linking mmr vaccine and autism was fraudulent. *BMJ*, 342, 2011.

[82] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley Longman Publishing Co., Inc., 1st edition, 1989.

[83] M. Gong, L. Ma, Q. Zhang, and L. Jiao. Community detection in networks by using multiobjective evolutionary algorithm with decomposition. *Physica A: Statistical Mechanics and its Applications*, 391(15):4050–4060, 2012.

[84] GovTrack.us. Senate voting data in 2014, 2014 (accessed May 3, 2016).

[85] S. Gregory. An algorithm to find overlapping community structure in networks. In *Knowledge discovery in databases: PKDD 2007*, pages 91–102. Springer, 2007.

[86] S. Gregory. A fast algorithm to find overlapping communities in networks. In *Machine learning and knowledge discovery in databases*, pages 408–423. Springer, 2008.

[87] S. Gregory. Fuzzy overlapping communities in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02017, 2011.

[88] D. M. Hartley, N. P. Nelson, R. Walters, R. Arthur, R. Yangarber, L. Madoff, J. Linge, A. Mawudeku, N. Collier, J. S. Brownstein, et al. The landscape of international event-based biosurveillance. *Emerging Health Threats*, 3, 2010.

[89] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4–6):175–181, 2000.

[90] J. H. Holland. *Adaptation in natural and artificial systems.* MIT Press, Cambridge, MA, USA, 1992.

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

[92] Y. hong Dong, Y. Zhuang, K. Chen, and X. Tai. A hierarchical clustering algorithm based on fuzzy graph connectedness. *Fuzzy Sets and Systems*, 157(13):1760–1774, 2006.

[93] E. Hruschka, R. Campello, A. Freitas, and A. de Carvalho. A survey of evolutionary algorithms for clustering. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 39(2):133 –155, march 2009.

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

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

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

[97] M. N. Kamel Boulos, A. P. Sanfilippo, C. D. Corley, and S. Wheeler. Social web mining and exploitation for serious applications: Technosocial predictive analytics and related technologies for public health, environmental and national security surveillance. *Computer Methods and Programs in Biomedicine*, 100(1):16–23, 2010.

[98] R. Kannan, S. Vempala, and A. Veta. On clusterings-good, bad and spectral. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, FOCS '00, pages 367–, Washington, DC, USA, 2000. IEEE Computer Society.

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

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

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

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

[103] M. Keller, M. Blench, H. Tolentino, C. C. Freifeld, K. D. Mandl, A. Mawudeku, G. Eysenbach, and J. S. Brownstein. Use of unstructured event-based reports for global infectious disease surveillance. *Emerging infectious diseases*, 15(5):689, 2009.

[104] M. Keller, C. C. Freifeld, and J. S. Brownstein. Automated vocabulary discovery for geo-parsing online epidemic intelligence. *BMC bioinformatics*, 10(1):385, 2009.

[105] Y. Kim and H. Jeong. Map equation for link communities. *Physical Review E*, 84(2):026110, 2011.

[106] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press, 1992.

[107] J. M. Kumpula, M. Kivelä, K. Kaski, and J. Saramäki. Sequential algorithm for fast clique percolation. *Physical Review E*, 78(2):026109, 2008.

[108] G. N. Lance and W. T. Williams. A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems. *The Computer Journal*, 9(4):373–380, Feb. 1967.

[109] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[110] W. Langdon and R. Poli. Evolving problems to learn about particle swarm and other optimisers. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, volume 1, pages 81 –88 Vol.1, sept. 2005.

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

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

[113] 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):606–613, 2013.

[114] K. Lee, A. Agrawal, and A. Choudhary. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1474–1477. ACM, 2013.

[115] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM, 2007.

[116] Y. Li, J. Chen, R. Liu, and J. Wu. A spectral clustering-based adaptive hybrid multi-objective harmony search algorithm for community detection. In *Evolutionary Computation (CEC), 2012 IEEE Congress on*, pages 1–8. IEEE, 2012.

[117] J. P. Linge, J. Belyaeva, R. Steinberger, M. Gemo, F. Fuart, D. Al-Khudhairy, S. Bucci, R. Yangarber, and E. van der Goot. Medisys: Medical information system. *Advanced ICTs for Disaster Management and Threat Detection: Collaborative and Distributed Frameworks*, pages 131–142, 2010.

[118] J. Liu, W. Zhong, H. A. Abbass, and D. G. Green. Separated and overlapping community detection in complex networks using multiobjective evolutionary algorithms. In *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pages 1–7. IEEE, 2010.

[119] Luciano, F. A. Rodrigues, G. Travieso, and V. P. R. Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, Aug. 2006.

[120] J. B. Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[121] A. Martín, H. D. Menéndez, and D. Camacho. Mocdroid: multi-objective evolutionary classifier for android malware detection. *Soft Computing*, pages 1–11, 2016.

[122] U. Maulik. Genetic algorithm-based clustering technique. *Pattern Recognition*, 33(9):1455–1465, 2000.

[123] H. Menéndez, D. F. Barrero, and D. Camacho. A multi-objective genetic graph-based clustering algorithm with memory optimization. In *2013 IEEE Conference on Evolutionary Computation*, volume 1, pages 3174–3181, June 20-23 2013.

[124] H. Menéndez and D. Camacho. A genetic graph-based clustering algorithm. In H. Yin, J. Costa, and G. Barreto, editors, *Intelligent Data Engineering and Automated Learning - IDEAL 2012*, volume 7435 of *Lecture Notes in Computer Science*, pages 216–225. Springer Berlin / Heidelberg, 2012.

[125] H. D. Menendez, D. F. Barrero, and D. Camacho. A co-evolutionary multi-objective approach for a k-adaptive graph-based clustering algorithm. In *Evolutionary Computation (CEC), 2014 IEEE Congress on*, pages 2724–2731. IEEE, 2014.

[126] E. Mykhalovskiy and L. Weir. The global public health intelligence network and early warning outbreak detection. *Canadian journal of public health*, 97(1), 2006.

[127] M. Naldi, S. Salcedo-Sanz, L. Carro-Calvo, L. Laura, A. Portilla-Figueras, and G. F. Italiano. A traffic-based evolutionary algorithm for network clustering. *Appl. Soft Comput.*, 13(11):4303–4319, 2013.

[128] M. C. V. Nascimento and A. C. P. L. F. Carvalho. A graph clustering algorithm based on a clustering coefficient for weighted graphs. *J. Braz. Comp. Soc.*, 17(1):19–29, 2011.

[129] M. Newman. Network data, 2013 (accessed May 3, 2016).

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

[131] M. E. Newman. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, 103(23):8577–8582, June 2006.

[132] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133+, June 2004.

[133] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review*, E 69(026113), 2004.

[134] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri. Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(03):P03024, 2009.

[135] A. Ochoa Ortíz, A. E. Muñoz Zavala, and A. Hernández Aguirre. A hybrid system using pso and data mining for determining the ranking of a new participant in eurovision. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, GECCO '08, pages 1713–1714, New York, NY, USA, 2008. ACM.

[136] D. J. Opel and S. B. Omer. Measles, mandates, and making vaccination the default option. *JAMA pediatrics*, 2015.

[137] L. Ott, M. Longnecker, and R. L. Ott. *An introduction to statistical methods and data analysis*, volume 511. Duxbury Pacific Grove, CA, 2001.

[138] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[139] C. Paquet, D. Coulombier, R. Kaiser, and M. Ciotti. Epidemic intelligence: a new framework for strengthening disease surveillance in europe. *Euro surveillance: bulletin europeen sur les maladies transmissibles= European communicable disease bulletin*, 11(12):212–214, 2005.

[140] M. Phillips. It's time to make our minds up on europe. *The Observer*, (Friday 12), March 2004.

[141] C. Pizzuti. Ga-net: A genetic algorithm for community detection in social networks. In *Parallel Problem Solving from Nature–PPSN X*, pages 1081–1090. Springer, 2008.

[142] C. Pizzuti. Overlapped community detection in complex networks. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pages 859–866. ACM, 2009.

[143] C. Pizzuti. A multiobjective genetic algorithm to find communities in complex networks. *Evolutionary Computation, IEEE Transactions on*, 16(3):418–430, 2012.

[144] R. Poli and W. B. Langdon. Backward-chaining evolutionary algorithms. *Artificial Intelligence*, 170(11):953 – 982, 2006.

[145] R. Poli, W. B. Langdon, and N. F. McPhee. *A Field Guide to Genetic Programming*. Lulu Enterprises, UK Ltd, 2008.

[146] P. Pons and M. Latapy. Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005*, pages 284–293. Springer, 2005.

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

[148] C. Ramirez-Atencia, G. Bello-Orgaz, M. D. R-Moreno, and D. Camacho. A Hybrid MOGA-CSP for Multi-UAV Mission Planning. In *Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference*, pages 1205–1208. ACM, 2015.

[149] J. Reichardt and S. Bornholdt. Detecting fuzzy community structures in complex networks with a potts model. *Physical Review Letters*, 93(21):218701, 2004.

[150] T. Richardson, P. J. Mucha, and M. A. Porter. Spectral tripartitioning of networks. *Physical Review E*, 80(3):036111, 2009.

[151] J. Ritterman, M. Osborne, and E. Klein. Using prediction markets and twitter to predict a swine flu pandemic. In *1st international workshop on mining social media*, 2009.

[152] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

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

[154] F. Santo. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[155] K. Sarma and H. Adeli. Fuzzy genetic algorithm for optimization of steel structures. *Journal of Structural Engineering*, 126(5):596–604, 2000.

[156] S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[157] P. Schuetz and A. Caflisch. Multistep greedy algorithm identifies community structure in real-world and computer-generated networks. *Physical Review E*, 78(2):026112, 2008.

[158] H.-P. Schwefel. *Numerical Optimization of Computer Models.* John Wiley & Sons, Inc., New York, NY, USA, 1981.

[159] J. Scott. *Social network analysis.* Sage, 2012.

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

[161] R. Shang, J. Bai, L. Jiao, and C. Jin. Community detection based on modularity and an improved genetic algorithm. *Physica A: Statistical Mechanics and its Applications*, 392(5):1215–1231, 2013.

[162] L. Shaw, W. Spears, L. Billings, and P. Maxim. Effective vaccination policies. *Information Sciences*, 180(19):3728 – 3744, 2010.

[163] C. Shi, Y. Cai, D. Fu, Y. Dong, and B. Wu. A link clustering based overlapping community detection algorithm. *Data & Knowledge Engineering*, 87:394–404, 2013.

[164] L.-D. Shi, Y.-H. Shi, Y. Gao, L. Shang, and Y.-B. Yang. Xcsc:: A novel approach to clustering with extended classifier system. *International Journal of Neural Systems*, 21(1):79 – 93, 2011.

[165] M. Srinivas and L. Patnaik. Adaptive probabilities of crossover and mutation in genetic algorithms. *Systems, Man and Cybernetics, IEEE Transactions on*, 24(4):656 –667, apr 1994.

[166] N. Sunday. The online health care revolution: How the web helps americans take better care of themselves. *Pew Internet & American Life Project*, 2000.

[167] E. Tomita, A. Tanaka, and H. Takahashi. The worst-case time complexity for generating all maximal cliques and computational experiments. *Theor. Comput. Sci.*, 363(1):28–42, Oct. 2006.

[168] L. Y. Tseng and S. B. Yang. A genetic approach to the automatic clustering problem. *Pattern Recognition*, 34(2):415 – 424, 2001.

[169] 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):217, 2012.

[170] 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):637–641, 1998.

[171] D. Walk. 2013-2014 nba schedule, 2013 (accessed May 3, 2016).

[172] G. Wang, Y. Shen, and M. Ouyang. A vector partitioning approach to detecting community structure in complex networks. *Computers & Mathematics with Applications*, 55(12):2746–2752, 2008.

[173] T. Washio and H. Motoda. State of the art of graph-based data mining. *SIGKDD Explor. Newsl.*, 5(1):59–68, July 2003.

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

[175] D. J. Watts. *Small worlds : the dynamics of networks between order and randomness.* 1999.

[176] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

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

[178] J. Xie, S. Kelley, and B. K. Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR)*, 45(4):43, 2013.

[179] G. Yair. Unite unite europe' the political and cultural structures of europe as reflected in the eurovision song contest. *Social Networks*, 17(2):147–161, 1995.

[180] G. Yair and D. Maman. The persistent structure of hegemony in the eurovision song contest. *Acta Sociologica*, 39(3):309–325, 1996.

[181] J. Yang and J. Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596. ACM, 2013.

[182] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, 2015.

[183] J. Yang, J. McAuley, and J. Leskovec. Detecting cohesive and 2-mode communities indirected and undirected networks. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 323–332. ACM, 2014.

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

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

[186] S. Zhang, R.-S. Wang, and X.-S. Zhang. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications*, 374(1):483–490, 2007.

[187] H. Zhou and R. Lipowsky. Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. In *Computational Science-ICCS 2004*, pages 1062–1069. Springer, 2004.

[188] E. Zitzler, M. Laumanns, and L. Thiele. SPEA2: Improving the strength pareto evolutionary algorithm. In *Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems*, pages 95–100, 2001.