UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

PH.D. THESIS

IN

INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

MULTIMEDIA SOCIAL NETWORKS

TUTOR: PROF. ANTONIO PICARIELLO

COORDINATOR: PROF. DANIELE RICCIO

XXX CICLO

SCUOLA POLITECNICA E DELLE SCIENZE DI BASE
DIPARTIMENTO DI INGEGNERIA ELETTRICA E TECNOLOGIE DELL'INFORMAZIONE

Multimedia Social Networks

Giancarlo Sperlì

Supervisor: **Prof. A.Picariello**

Department of Information Technology and Electrical Engineering

This dissertation is submitted for the degree of
Doctor of Philosophy

December 2017

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Giancarlo Sperlì
December 2017

List of Acronyms

The following acronyms are used throughout this text.

HDFS Hadoop File System

IC Independent Cascade model

IM Influence Maximization

LTI Linear Threshold model

MSN Multimedia Social Network

OSN Online Social Network

SNA Social Network Analysis

TM Trivalency Model

WC Weighted Cascade Model

Abstract

Nowadays, On-Line Social Networks represent an interactive platform to share – and very often interact with – heterogeneous content for different purposes (e.g to comment events and facts, express and share personal opinions on specific topics, and so on), allowing millions of individuals to create on-line profiles and communicate personal information.

In this dissertation, we define a novel data model for Multimedia Social Networks (MSNs), i.e. social networks that combine information on users – belonging to one or more social communities – with the multimedia content that is generated and used within the related environments. The proposed data model, inspired by hypergraph-based approaches, allows to represent in a simple way all the different kinds of relationships that are typical of these environments (among multimedia contents, among users and multimedia content and among users themselves) and to enable several kinds of analytics and applications.

Exploiting the feature of MSN model, the following two main challenging problems have been addressed: the *Influence Maximization* and the *Community Detection*. Regarding the first problem, a novel influence diffusion model has been proposed that, learning recurrent user behaviors

from past logs, estimates the probability that a given user can influence the other ones, basically exploiting user to content actions. On the top of this model, several algorithms (based on game theory, epidemiological etc.) have been developed to address the Influence Maximization problem. Concerning the second challenge, we propose an algorithm that leverages both user interactions and multimedia content in terms of high and low-level features for identifying communities in heterogeneous network.

Finally, experimental analysis have been made on a real Multimedia Social Network ("Flickr") for evaluating both the feasibility of the model and the effectiveness of the proposed approaches for Influence Maximization and community detection.

Table of contents

Summary

In the last decade, our society has been transformed by the Internet and On-line Social Networks (OSNs) that connect people each other. In this type of society the connections between people assume an increasingly fundamental role because they let the flow of ideas to be exchanged among people. The study of how this flow of ideas changes over time is important because this is what drives humans' behaviors change and innovation. The study of how the flow of ideas led to make changes in behavior, is called *Social Physics* [106].

OSNs are really important because they represents a common means that people use for decision making; for instances, the adoption of a new technology or a new product is justified also by choosing of their social ties (e.g. colleagues, families, friends etc.), they feel more comfortable and this could mean in an increasing number of adopter of new technology. Thus, an increasingly important role is assumed by *Social Big Data*, that is defined as the set of process and methodologies to produce sensitive and relevant knowledge for any user or company from social media data sources.

The aim of this dissertetion is to define a novel data model for Multimedia Social Networks (MSNs) that combines information on users belonging to one or more social communities together with the multimedia content that is generated and used within the related environments. In particular, the proposed model relies on the hypergraph data structure to capture and to represent in a simple way all the different kinds of relationships that are typical of social networks and multimedia sharing systems, and in particular between multimedia contents, among users and multimedia content and among users themselves.

Chapter 1 introduces the concept of Big Data Analytics and Online Social Network analysis.

Chapter 2 discusses theoretical background about Social Networks and two main challenges facing in MSNs: *Influence analysis* and *Community detection*.

Chapter 3 focuses on the novel data model relying on hypergraph data structure to support in a simple way all the different kinds of relationships that are typical of Multimedia Social Network (MSN). This model leverages the intrinsic characteristic of multimedia with the aim to support several Big Data applications (such as influence analysis, lurker identification, expert finding and so on.).

Chapter 4 is composed by two sections. Firstly, it is described a novel approach for modeling the influence spread on MSNs based on activity logs containing actions that could be performed by users in one or more MSNs. On top of this model several algorithms have been developed to deal with Influence Maximization problem. In the second section an algorithm that exploits the feature of the MSN model is proposed to deal with the Community detection problem.

Chapter 5 describes our Big Data platform based on Spark technology and, in particular, the *SIMONA* layer developed to properly manage the Multimedia Social Network model. This infrastructure is used to

support the evaluation made on Flickr to evaluate the effectiveness and efficacy of the proposed approaches.

Chapter 6 provides the state of the art in literature about problems related the two main challenges over social networks discussed in Chapter 2. Eventually, a comparison between the proposed approaches and the state of the art algorithms is provided to analyze the novelties of the proposed techniques.

1 Introduction

In the last two decades, the amount of available data has grown rapidly. In 1992, the global Internet traffic was approximately 100 GB per day, today has reached more than 20,000 GB per second opening the Zettabyte [1] Era. Moreover, Cisco [2] predicts that:

- Annual global IP traffic will reach 3.3 ZB per year by 2021
- Every second, a million minutes of video content will cross the network by 2021
- Global Internet traffic in 2021 will be equivalent to 135x the volume of the entire Global Internet in 2005

The data explosion has been fueled specially by the digital revolution and the rapid advances in technology. More and more devices are used to capture and share the data such as personal devices (mobile phones, personal computers) and sensors IoT [3]. These has changed the types of

[1] 1 ZB = 1 trilliongigabytes.

[2] https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf

[3] Internet of Things

data that can be collected and shared, indeed, the recent smartphones allow to capture text, audio, video, and GPS data. Nowadays, it is possible to note that smartphone traffic exceeds PC traffic and it expects that there will be 27.1 billion networked devices in 2021

This growth trend involves several sectors and consequently the analysis of data assumes an important role for many companies. On one hand it represents the lifeblood for this company on other hand managing these information becomes more difficult. To gather knowledge from this raw materia requires a new approach that involves the Big Data and related technologies.

1.1 Big Data

There are several definitions of Big Data. It is possible to refer to it through some characteristics and properties concerning to the volume of data or the richness of data, as Gartner [4] says about it: "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making".

To better understand the *Big Data* ecosystem, in a *2001 MetaGroup publication*, *Gartner analyst Doug Laney* introduced the 3 V's of data management, which allow to define the 3 main components of data as *Volume, Velocity and Variety*:

- Volume refers to the vast amounts of data generated every second, for example all the emails, twitter messages, photos, video clips, sensor data...To operate, store and use this amount of data, distributed systems are used, where parts of the data is stored in different locations and brought together by software;

[4] https://www.gartner.com/it-glossary/big-data/

- Velocity refers to the speed at which new data is generated and the speed at which data moves around. Tecnology allows us to reacting fast enough and analyzing the streaming data while it is begin generated, without ever putting it into databases;

- Variety refers to the different forms of data that we collect and use. In the past, the data was strictly structured and their management was carried out by using tables or relational databases, now data comes in different formats, such as structured and unstructured increasing its complexity that excludes the traditional means of analysis.

Others definitions involve additional Vs such as *Veracity*, introduced by *IBM*:

- Veracity refers to the uncertainty of data (biases, noise and abnormality), that makes data inconsistent and incomplete, which leads to another challenge, keeping big data organized.

The term Big Data is also referred to data that is difficult to process using traditional relational database systems. Unstructured and semi-structured data types typically don't fit well in traditional databases, where is requested to transform the data before to save it in tables. To manage far larger quantities of data than before is required massively parallel-processing (MPP) running on clusters of thousands of commodity machines. This facet requires a new processing technologies like Google's MapReduce or its open-source equivalent, Hadoop.

1.2 Big Data Analytics

Big data analytics is the process of examining large and varied data sets that involves data scientists, predictive modelers, and statisticians

to analyze and to transforms data in knowledge and subsequently in business decisions.

Big data analysis has an important role in recent companies that incorporate them into processes of Business Intelligence (BI). Big Data can be analyzed for insights analysis leading to assume business and strategic decisions. This process of analysis is based on extract, transform and load (ETL), it also includes tools for data mining, predictive analytics and machine learning.
A *non-relational database* is any database that does not follow the relational model provided by traditional relational database management systems. This category of databases, has been increasingly used in recent years with the rise of *Big Data Application*, because *NoSQL* can incorporate literally any type of data, while providing all the features needed to build content-rich apps. *NoSQL* databases and management systems are relation-less (or *schema-less*), and they are not based on a single model (*e.g. relational model of RDBMSs*) and each database, depending on their target-functionality, adopt a different one. This approach is used for:

- *Simplicity of design* unlike relational models do not require a default schema, *NoSQL* databases offer flexible schema design that make it much easier to update the database to handle changing application requirements;

- *data structure* non relational databases are designed to handle unstructured data;

- *"horizontal"scaling to cluster of machines* (not allowed from relational databases), by taking advantage of cheap, commodity servers;

- *finer control over availability*;

- *Open source* which means you don't have to pay any software licensing fees upfront.

There are four general types of *NoSQL* databases, each with their own specific attributes:

- Graph database: based on graph theory, these databases are designed for data whose relations are well represented as a graph and they have elements which are interconnected, with an undetermined number of relations between them;

- Key-Value store: these databases are designed for storing data in a *schema-less* way. In a *key-value store*, all of the data within consists of an indexed key and a value, hence the name;

- Column store: instead of storing data in rows, these databases are designed for storing data tables as sections of columns of data, rather than as rows of data, for these reason they offer very high performance and a highly scalable architecture;

- Document database: expands on the basic idea of key-value stores where "documents" contain complex data and each document is assigned a unique key, which is used to retrieve the document. These are designed for storing, retrieving, and managing document-oriented information, also known as semi-structured data.

The tables 1.1, 1.2 lays out some of the key attributes that should be considered when evaluating *NoSQL databases*:

Datamodel	***Performance***	***Scalability***	***Flexibility***
Key-Value Store	High	High	High
Column Store	High	High	Moderate
Document Store	High	Variable(High)	High
Graph Database	Variable	Variable	High

Table 1.1: NoSQL attributes

Data model	***Complexity***	***Functionality***
Key-Value Store	None	Variable(None)
Column Store	Low	Minimal
Document Store	Low	Variable(Low)
Graph Database	High	Graph Theory

Table 1.2: NoSQL attributes

NoSQL database environments are built with a distributed architecture so there are no single points of failure and there is built-in redundancy of both function and data, so that if one or more database servers, or *'nodes'* goes down, the other nodes in the system are able to continue with the operations without data loss, showing true fault tolerance. In the Internet age the concept of transaction it is changed and it has been demonstrated that *ACID* transactions are no longer a requirement in database driven systems. The *"C"* in *ACID* refers to *data Consistency* in relational database management systems which is enforced via foreign keys/referential integrity constraints. This type of consistency is not utilized in *NoSQL* databases because there are no *JOIN* operations. The *Consistency* that concerns *NoSQL* databases is found in the *CAP* theorem, which signifies the immediate or eventual consistency of data across all nodes that participate in a distributed database.

1.3 Social Networks

A social network is formed by a whole or multiple sets of individuals who share social relationships between them as casual knowledge, friendship, family or sentimental ties. The subject of the analysis is the individual. Instead, contacts, constraints, and links between two or more actors are relational data. Relational structures are a relevant social form that defines the context in which the actors move. The relationship's properties describe actor's behavior within the network. Therefore, the analysis has the dual task of describing the network and predicting it and explaining its evolution. Knowledge of actors, relationships and their attributes, which can represent behaviors and opinions of individuals and groups, allows to define one or more "maps" of the social network.

In recent years with the advent of the Internet, Web 2.0 and the Online Social Network (OSN) the expression "social network" has become a common vocabulary. Their introduction has allowed people living in different parts of the world to create relationships of different kinds and to share, comment and observe various types of multimedia content. OSN is often associated with the more general concept of "social media": it refers to online technologies and procedures that social network users use to share textual contents, images, videos, and audios. In addition, we may also consider Multimedia Social Network (MSN)[2], i.e. a network that combines the users information belonging to one or more social communities to the multimedia content generated by them. Note that in "interest-based" networks, each user interacts with others through these contents, encouraging relationships that can help them understand their behavior within the network.

The 31% of internet traffic is generated by on-line social networks[5]:

[5]https://www.webpagefx.com/internet-real-time

- Every day, Facebook users like an average of 4.5 billion posts, share more than 4.7 billion status updates with their friends or followers, and watch over 1 billion videos.

- Instagram, which recently reached the 300 million monthly active user mark, continues to grow at breakneck speed. Presently, the app sees 2.5 billion likes daily – more than 28,000 every second – and 70 million new photo uploads per day.

- LinkedIn, the social media site for professionals, has become a hotbed of activity for those looking for coworkers to connect with, recommendations to add to their profile, new career opportunities, and now even content to read and share. In 2012, professionals performed more than 5.7 billion searches on LinkedIn, which breaks down to more than 180 searches per second.

- In the 10 years since its inception, Flickr has grown its photo sharing platform to boast over 92 million users and 2 million groups. Approximately a million photos are uploaded to the site every day, which was acquired by Yahoo! in 2005.

- According to YouTube, more than 1 billion unique users visit the site each month and consume over 6 billion hours of video – almost an hour for every person on Earth. Every second, approximately 2,314 hours worth of video is consumed. Additionally, 100 hours of new video is uploaded to YouTube each minute.

- Twitter's average TPS – or tweets per second – is in the ballpark of 5,700. The last time a major spike to the TPS was experienced was in August of 2013, when Japanese users watching Studio Ghibli's "Castle in the Sky" spiked the activity to a record 143,199 tweets each second. Not even the 2014 World Cup was able to beat this per-second record: the final moments of the Germany-Brazil

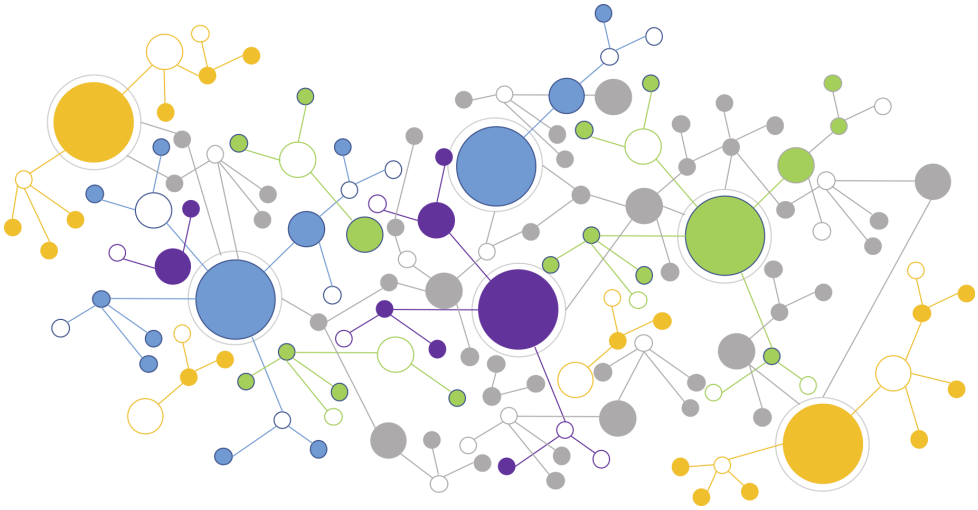


Figure 1.1: An example of Social network

semifinal clocked in at only 9,669 TPS (but it did set a record for most tweets per minute).

OSN may be thus considered as one of the main Big Data sources, considering the usual very high quantity of managed data (Volue), the frequencey of data changes (Velocity) and the huge heterogeneity of multimedia data (Variety). This is why to manage this kind of information represents a very good opportunity for Big Data Analytics.

1.4 Social Networks Analysis

Social networks and OSNs are becoming very important for analyzing interactions among people. In a nutshell, it allows the "study of human relationships by means of graph theory"[140].

OSNs such as Facebook, Twitter, LinkedIn and Multimedia Networks as Flickr and Instagram, have become very popular over the last two decades, justified by the spread of internet enabled devices such as smartphones, tablets, personal computers and other devices.

Social networks like Facebook are designed for social interaction, while others such as Flickr are designed for content sharing services, allowing people to improve their social connections.

They represent a new way of meeting people, introducing new forms of relationships and interaction that did not exist some years ago. In addition, OSNs provide a new opportunity to analyze social interactions on larger scale.

Social Networks contain the information about the links that represent the structure of the graph in which the different entities interact. We can represent a social network through a graph where nodes are made up of individual actors and multimedia objects, and the edges are made up of the interactions or relationships between these entities.

The study of social networks began in the 20th century with the aim to identify, measure and represent social relationships between individuals, groups, organizations and other actors involved in the exchange of information and knowledge. It thus allows to determine the influence of the social relationship on collective behavior. All Applications involving social network analysis are multiple. The analyzes follow three fundamental points:

1. Focusing on the properties of social networks in the real world;

2. Considering the social network as something evolving following a set of rules;

3. Studying social networks by following not only topology but also collective behavior.

Social Networks Analysis (SNA) involves several kinds of data analytic application such as data mining, machine learning, text analysis, image analysis, in this last case is more correct to talk about Multimedia Social Networks Analysis to emphasize the presence of the media. Through

the media it is possible identify a rich variety of interaction between the actors of the social networks.

OSNs have led to a large increase of information about individuals and their interaction. As a result, the major challenge is make more efficient the data collection and the analysis of the big volume of data, characterized by big variety and big heterogeneous. These characteristic places the OSNs Analysis in the context of the Big Data Analytics.

The ability of treat this rich amount of content is an opportunity to improve the knowledge and allow to use this information in a wide variety of fields such as Social science and Marketing:

- Recommendation System to suggest products based on the study of users' interest.
- Sociology study of communities, such as the research of an interest within a community.
- Suggesting new contacts and interactions between users.
- Advertising System and Influence maximization with which is possible to improve the number of individuals reached.
- Influence Analysis to find the most influencer users.

The social networks are often analyzed according to two approaches:

- Linkage-based and Structural Analysis: In this case we exploit the network's structure to obtain information about the important nodes, communities and links.
- Adding Content-based Analysis: nowadays there are a lot of multimedia social networks characterized by a large amount of content, these data can be use to improve the quality of the analysis.

2 Theoretical Background

2.1 Social Network Theory

For long time social philosophers and scientists were interested in the reasons that lead people to establish social relations. This interest is particularly focused on the analysis of relationships between social entities and their meaning. In the early 1930's Moreno [96] proposed a "social configuration", called *sociogram*, marking the beginning of sociometry, that is the precursor of Social Network Analysis. A *sociogram* is a structure to represent social entities and their relationships (i.e. friendships, interpersonal choice etc.) respectively as points in two dimensional space and lines linking the corresponding points. These structures can be used by researchers to identify leaders and groups of people and to represent connection chain. Lewin [84] analyze the group behaviour, outlining that a social group is determined by social forces defined in a social 'space' which includes the group together with its surrounding environment. Moreover, Lewin [85] stated that mathematical techniques of topology and set theory can be used to study the structural properties of this social space.

Anthropologists focused their attention on more complex models to overcome the limits of the traditional approach of describing social organization in terms of institutions (economics, religion, politics, kinship, etc.) for understanding the behavior of individuals in complex societies.

In turn, in psychological field, many researchers studied group process, describing actors as points and channel of communications as lines among them. Other researches analyzed the organizations, that is groups of people, as social groups with stable patterns of interaction over time [74]. The interest in the analysis of social relations is due to the concept that "Social life is relational", having pointed out by Collins [32]. The fundamental elements in this approach is the *social atom*, that is an entity that interacts with others based on social behaviors. A relation can be seen as a linkage or a flow among these entities that can assume several meanings such as acquaintance, kinship, physical connections and so on. A particular structure based on social atoms and relations is called "Social Network", term used the first time by Barnes[16]. The aim of Barnes was to study the conncetions among people living in a Norweian island parish, describing them respectively as a set of points interconnecting by a set of lines that represent the relations among people. A first definition of social network was proposed by Mitchel [94] that defines it as "a specific set of linkages among a defined set of persons, with the additional property that the characteristics of these linkages as a whole may be used to interpret the social behavior of the persons involved". This definition views organizations in society as a system of objects (e.g. people, groups, organizations) joined by a variety of relationships. Some pairs of objects are joined by multiple relationships. Wasserman and Galaskiewicz [53] provide another definition of Social Network, that is a network composed by a set of actors, particular social entities (individual, corporate or collective social atoms), and relations defined on them. In the provided definition, a volition or ability to ''act"

is not necessarly related to actor term. Moreover, Ellison [44] defines a Social Network as web-based services that allow individuals to construct a public or semi-public profile within a bounded system, articulate a list of other users with whom they share a connection, and view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site.

In the 1960, Stanley Milgram performed an experiment to find out the average path between two people in United Stated. This experiment showed that Social Networks have the property of Small World Network, particular network in which the distance, corresponding to the number of steps, between two nodes grows proportionally to the logarithm of the size of vertices set. The experimental procedure expected that he would send hundreds of letters to people in Nebraska and that they would reply directly in case of they knew the sender or forward correspondence to someone to reach the designated target person living in Boston as quick as possible. This experiment had as outcome that human society is a small-world network characterized by shortest path-length equals to six, the phenomenon colloquially known as the six degrees of separation.

Granovetter [61] analyzed the job seeking in the 1960s and 1970s for understanding the usefulness of different types of ties in certain situation. The result of this analysis was that a social network could be partitioned into "strong" and "weak" ties.

In Social Networks' field another main property is the *triadic closure*. This property is related to the increase of likelihood that two u_1 and u_2 users, who have a common friend u_3, will become friends in the future. The term "triadic closure" stems from the consideration that the future friendship between the two u_1 and u_2 users allows to close the triangle with u_3. This principle can explain the evolving of network over times in many situations based on the following three reasons: (i) Opportunity,

based on the likelihood that u_1 and u_2 meet each other because they are both friends of u_3 (ii) Trusting, the common friendship with u_3 gives them a reason to trust each other (iii) Incentive, a source of latent stress in these relationships if u_1 and u_2 are not friends with each other.

Social Network Analysis (SNA) is a set of techniques and methodologies to the properties of Social Network with the aim of supporting a wide range of applications. SNA have been widely used in different fields, in which entities and relationships assume different meaning based on the examined application. They can be used to represent the personal contacts across organizational boundaries [60], the spread of ideas, information and technlogies [116] or to analyze the recruitement process of members of social movement organization by friends or acquaintances [128]. In management consulting, network analysis is often applied in the context of knowledge management, where the objective is to help organizations better exploit the knowledge and capabilities distributed across its members. In public health, network approaches have been important both in stopping the spread of infectious diseases and in providing better health care and social support.

Thus, a social network is a social structure made up of individuals (or organizations) called "nodes", which are tied (connected) by one or more specific types of interdependency, such as friendship, kinship, common interest, financial exchange, dislike, sexual relationships, or relationships of beliefs, knowledge or prestige. Thus, the Social Network model relies on the Entity concept, that can be simply preliminarly defined as a collection of occurrences having same properties. In other words, in order to describe the generic interaction among entities, it is possible to define the relationship concept as a conceptual link between two or more entities involved in a Social Network.

Definition 2.1.1 (Relationship - Preliminary Definition)**.** *Let $E_1, E_2, \ldots, E_n$ be n entities, a relationship R is defined as:*

$$R \subseteq E_1 \times E_2 \times \ldots \times E_n$$

where each entity E_i is involved in relationship R as well as the occurrences of entities e_i is an element of single relationship instance $r_i = (e_1, e_2, \ldots, e_n)$.

The relationships can be described by a set of attributes that can be also used to identify univocally a given occurrence; for instances timestamp, sentiment, text etc. The cardinality of relationship is the number of entities involved in the relationship itself.

We classify the different relationships that can be instantiated in the network in two families:

- direct relationship corresponding to action made by users; such as friendship, tagging, endorsement and so on.
- indirect relationship corresponding to similarity connection among entities of the same type.

Thus, it is possible to provide the following preliminary definition of a general Social Network model.

Definition 2.1.2 (Social Network - Preliminary Definition)**.** *Given that a set of data D, which correspond to a collection of information belonging to an On-line Social Network, a Social Network is composed by a set of entities $\mathcal{E}$, a set of relationships R.*

2.2 On-line Social Network

In the last decades, the growth of On-line Social Networks (OSNs) has led to the development of a new communication means that allows people

around world to interact among them and spread information in the network using Internet technologies. As common in the Web 2.0, there is not a generally one well-established definiton for OSNs. Schneider et al. [123] define OSNs as a set of communities composed by people that share common interest, activities, backgrounds, and/or friendships and can interact with others in numerous ways. Boyd and Ellison [44] define a OSN as a social networking site (SNS) that allows users to create a public or semi-public profile, a users list with whom share information and browse its connection list or those made by others within the system. The provided definition does not contain the term "social networking" because it emphatizes the relationship concept, that often involve strangers. Other definitions consider OSNs as a particular type of virtual community [43] and of social software [114]. Spagnoletti et al. [130] describe Social Media (e.g. Twitter, Instagram, Facebook etc.) as IS artefact composed by three systems: (i) technological, used to solve a problem, achieve a goal or serve a purpose defined by humans support or hinder social interactions; (ii) informational, the instantiation of information occurring through a human act consists of user generated digital content; and (iii) social, the nature of the relationships or interactions between individuals (hence social) in attempts to solve problems, achieve goals or serve ones purposes. communication and collaboration activities.

Thus, the study of human relationships has been revoluzionated by the growth of On line Social Network. In this context, social links allows to build, for example, networks of professionals and contacts (e.g., LinkedIn, Facebook, MySpace) and networks for sharing content (e.g., Flickr, YouTube). Several studies have been made to identify the users' motives for using OSNs [43]. As underlying of vom Brocke et al. [23], one of the main motive is the "identity" management, corresponding to the constructing and maintaining of a personal profile to present oneself to other users.

The foundation of social identity theory was postulated by Tajfel [132], connecting the following three social-psychological processes:

- Social categorization: The perception of people to be component of a group;
- Social comparison: the tendency to evaluate the worth of groups as well as individuals to compare with other groups;
- Social identification: The notion that people are conditioned by perception and responses to social situations.

The interest of psychologists has been attracted by social influence processes, and in particular how they can influence the behavior or opinions of users over time. Golbeck et al. [56] underlyine the importance to analyze the people's behavior in On-line Social Network to better understand their personality for two reasons. Firstly, OSNs provide platform allowing users free interaction and exposing viewpoints to satisfy users' basic psychological needs. Secondly, there is an ample amount of data regarding normative behaviors of individuals which guarantees fair analysis of individual's personality. Moreover, Asendorpf and Wilpers [11] established that the individuals' propensity to positive and negative relations is determined by people personality.

In OSNs, users are able to present themselves by user profiles, that contains both personal information and personal interests. A interesting study is provided by Ross et al.[118] with the aim to analyze the correlation between the personality of the individual user and their behavior on a social network. This study is based on the self-reports of users of Facebook and measured five personality factors using the NEO-PI-R [36] questionnaire for inferring connections between the personalities of surfers and their behavior on Facebook. The Big-Five model, proposed by Costa and McCrae [37], is composed by the following five traits for

describing people personality: Extraversion, Openness to experience, Agreeableness and Neuroticism.

Online Social Networks (OSNs) allow users to have an online presence defining a user profile that contains shareable personal information, such as a birthday, hobbies, preferences, photographs, writings, etc. Most SNSs offer features of convenience that help users form and maintain an online network with other users. An example is the friendship relationship, representing a shortcut to others' front pages and, in some SNSs, a convenient tool to exchange messages and stay in touch. They can add friends, invite new friends, join groups or networks, communicate with other users via OSN-internal email services, writing on other users' "walls" or on discussion forums, and adjust their privacy settings.

Most OSNs include features for creating user accounts and authenticating users. A user's basic profile contains home town, age and other personal information. Several features are provided by OSNs to users that are logged in; in particular, they can update their profile(i.e., personal information, photograph etc.) or browse other users' profiles. OSNs offer a variety of different features that are commonly accessible only to those users that are logged in. Users can update their profiles (contact information, photographs, information about hobbies, books, movies, music, etc.), browse other users' profiles by searching and subsequently obtaining lists of their friends and narrowing them via categories like schools or work sites.

Several studies have been made to analyze user behaviors by the analysis of interaction events' records on different links. Recently, two significant studies [18] [123] used clickstream data at the network level to capture the behavior of OSN users and revealed that passive or latent interactions, such as profile browsing, often dominate user events in a social network. Other studies have the aim to analyze the user behavior for different different purpose. Firstly, site design [152][25] and adver-

tisement placement policies [88] can be improved by the understanding of user behavior on the network. In OSNs, the spread of contents or promotions [115] [82] can be improved by a better understanding of user behaviors. Moreover, multimedia object (e.g. photo, video, status etc.) can be used to improve the spread information [159]. Eventually, Internet infrastructure and content distributions systems [109] [78] can be improved by the analysis of workload of social networks.

In an OSN a user profile is described by a list of interests based on its' taste preferences. Lui [91] analyzed these "lists of interests" in user profiles determining that it represents an idealized projection of oneself.

2.2.1 Classification

In the last decade the continuous increase of users in OSNs and the technological development allowed to define new ways to interact among people in which multimedia contents assume a key role, calling them as Social Media Networks. These facets led to emerge and arise of new challenges and new research topics. In particular, these challenges concern recommendation applications (such as friend and media recommendation, recommendation for influence maximization, location and itinerary recommendation and so on), Marketing applications (Influence Maximization, Influence diffusion, Viral Marketing, Word-of-mouth and so on) and Security field (Lurker detection, malicious user detection and so on).

Social Media Networks can be classified in the following three categories:

1. OSNs: In this type of networks it is possible to exploit the well-known relationships established among users, users and contents for different purpose, such as identify the most appropriate content for specific users, ranking of social network users respect to a given item, identify a subset of user to maximize the influence of specific

contents and so on. In this categories we can identify the following social networks: Twitter, Facebook, Instagram, Google+, Last.FM, Flickr.

2. Location Based Social Networks (LBSNs): These networks introduce location as a new object, that are used for defining user's context and behaviors analyzing the relationships established among users, location and users and locations. It is possible to identify in this class the following social networks: Foursquare, Yelp, TripAdvisor.

3. Social Movie Rating Businesses (SMRBs): These networks introduce movie as the main entity that users can rate, review or comment. Examples of this category are Flixster, IMDB, Movielens.

4. Social Content Delivery Networks: these networks leverage the content delivery paradigm to share multimedia contents with users that are interested in. Examples of this category are Youtube, Spotify and so on.

2.3 Multimedia Social Network

The development of technology has enhanced OSN features, enabling users to share their life by describing multimedia content (text, audio, video, images) and interacting with such objects in order to provide feedback or comments or feelings compared to them. In the last years, several OSNs (such as Facebook, Twitter, and so on) provide several features to share and interact with multimedia contents (audio, video, images, texts). This led to the definition of a new type of OSN, called *Multimedia Social Network* (MSN), in which users share and exchange multimedia contents (such as music, images, posts etc).

In the literature, the term MSN have been used over the last years together with *Social Multimedia Network* or *Social Media Network* to indicate information networks that leverage multimedia data in a social environment for different purposes: distributed resource allocation for multimedia content sharing in cloud-based systems [99], generation of personalized multimedia information recommendations in response to specific targets of interests [90], evaluation of the trust relationship among users [161], high dimensional video data distribution in social multimedia applications [68], characterization of user behavior and information propagation on the base of multimedia sharing [102], representation of a social collaboration network of archeologists for cultural heritage applications [98], just to cite some of the most recent proposals.

In MSN multimedia contents play an increasingly important role, which can be used by integrating them with user interactions with them to support different applications such as influence analysis community detection viral marketing. For instance, comments on Youtube are used by De Choudhury et al. [39] to infer users' "interests" and topic. Moreover, chat activity in Instant Messenger [126] and content and trend analysis of Twitter messages [125] are used to analyze respectively the multimedia contents about video and events.

In such a context, *Multimedia Social Networks* actually represent a natural environment where users can create and share multimedia content such as text, image, video, audio, and so on. Just as an example, each minute thousands of tweets are sent on Twitter, several hundreds of hours of videos are uploaded to YouTube, and a huge quantity of photos are shared on Instagram or uploaded to Flickr.

Indeed, using multimedia content each user interacts with the others generating "social links" that well characterize the behaviors of users in the network. In particular, the links in a social media network could represent anything from intimate friendships to common interests for

a given multimedia object (e.g., tweet, post, video, photo, etc.): they determine the "flow" of information and hence indicate a user's *influence* on the others, a concept that is crucial in sociology and viral marketing.

Thus, in a *Multimedia Social Network* multimedia data can play a "key-role" in the SNA: representing and understanding all the possible "user-to-multimedia" interaction mechanisms can be useful to better predict user behavior and estimate the related influence in the network. In this case, additional research questions have to be addressed under the *Social Network Analysis* (SNA) perspective:

- Is it possible to exploit multimedia features and the notion of similarity among multimedia contents to discover more useful social links?
- Can all the different types of user annotations (e.g. tag, comment, review, etc.)[6] and interactions with multimedia objects provide a further support for an advanced network analysis?
- Is it possible to integrate and efficiently manage in a unique network[5] the information coming from heterogeneous social media networks (for example, an Instagram user has usually an account also on Flickr)?

To capture the described issues, the term *Multimedia Social Networks* (MSNs) is defined as an "*integrated social media networks that combine the information on users, belonging to one or more social communities, together with all the multimedia contents that can be generated and used within the related environments*".

Tian [139] provides a classification of MSN into the following three categories:

1. Imagery Social Networks: these networks allows to capture social and activity relationships between users through multimedia

contents captured by surveillance videos, or wireless sensors. In these networks, data privacy is one of the main problem concerning user's location or other private information. An analysis of background factors, motivations and social network experience of users is provided by Litt [89] to understand how these factors can be exploited to protect their data.

2. Gaming-Driven Social Networks: In these networks users share and interact with multimedia contents to maximize their profile, that also analyze other users behaviors to achieve effective cooperation. Peer-to-peer (P2P) network is a representative example of this category.

3. Interaction-Driven Social Networks: In these networks the relationships represents users' interaction and other activity made by users in OSNs, that could be used to develop On-line Social multimedia services.

2.4 Challenges in Multimedia Social Network

The use of OSNs is rapidly growing allowing people, living different places, to make friends and to share, comment and observe different types of content. OSNs have thus produced a tremendous amount of data showing Big Data features, which led to the growth of new challenges.

In a Social Network people are often affected by the behavior of their neighbors, that could be their friends, colleagues, people belonging to same social ties/communities etc.. The structure of a social network is commonly represented by a graph, that consists of a set of nodes (or vertices) that represent individuals, groups and entities, and a set of links that connect a subset of possible pairs of vertices and represent the

relationships or interactions that exist between the single pairs. Formally, the Online social network is usually represented as weighted direct graph $G = (V, E, \omega)$, where each edge $e \in E$ is assigned a weight ($p_{u,v} \in [0, 1]$) that describe the strength of the interaction between two nodes.

In this section, two main problems about Multimedia Social Network are described.

2.4.1 Influence Analysis & Maximization

The emergence of new information and communication technologies, particularly the Internet and social networks, have changed consumer consumption habits by providing consumers with new ways of looking for, assessing, choosing, and buying goods and services [3], increasing the power of consumers [142]. Viral marketing capitalizes on the advantages of social networks including their high capacity for diffusion of information. In this sense there are a couple of laws and theories relating to the utility of networks and other aspects such as the critical mass of connectivity required for a network to be valuable.The effectiveness of social media marketing campaign depends on how users perceive the companies and brands they interact with. They can be perceived as "interlopers", “party crashers" [50], or unwanted guests in the interactive space [124]. The interaction among firms and customers led to build brand loyalty through the promotion of products and services as well as the setting up of online communities of brand followers [72]. Furthermore, the interaction among customers allows firms to increase the trend reputation exploiting new means and channel [63]. Moreover, social media marketing campaign is also influenced by the type of industry and product. For instance, in the hotel industry Corstjens and Umblijs [34] show how the firm reputation impacts the effectiveness of social media efforts. Thus, Social media improves the relationships among customers and brand improving exposure time and spread of marketing campaign

[141]. The influence process is a a process where ideas or behaviors are spread with the initiator and the recipient unaware of any intentional attempt at influence by giving advice and recommendations, by serving as a role model that others can imitate, by persuading or convincing others, or by way of contagion [147].

Rogers [117] provides a definition based on the communication concept describing as a process in which participants create and share information with one another in order to reach a mutual understanding. In particular, Rogers defines the Diffusion as the process in which an innovation is communicated through certain channels over time among the members of a social system. It is a special type of communication, in that the messages are concerned with new ideas. This definition implies that communication is a process of convergence (or divergence) as two or more individuals exchange information in order to move toward each other (or apart) in the meanings that they give to certain events. Diffusion is a special type of communication in which the messages are about a new idea. Moreover, Rice [Rice] provides another definition of diffusion in the context of media effects in which it is defined as "the process through which an innovation (an idea, product, technology, process, or service) spreads (more or less rapidly, in more or less the same form) through mass and digital media, and interpersonal and network communication, over time, through a social system, with a wide variety of consequences (positive and negative)". Rogers provides a classification of users in Social Networks in five adopter categories with the estimated percentage of people: 1. Innovators, 2.5 % 2. Early adopters, 13.5 % 3. Early majority, 34 % 4. Late majority, 34 % 5. Laggards, 16 %. Typically, people may change their belief based on the opinions of others that they feel similar with respect to psychological principles or the majority behaviors. Moreover, some people could be influenced by people that they believe experts in some fields. The social influence was postulated by French and

Raven [51] that define it as the outcome of the exertion of social power from one of five bases(reward power, coercive power, legitimate power, expert power or referent power). Moreover, it is possible to define *social influence* as change of an individual's thoughts, feelings, attitudes or behaviors that result from interaction with another individual or a group. Social influence has been defined as the process "wherein one person's attitudes, cognitions, or behaviors are changed through the doings of another" [17] and as "the maryiad ways that people impact one another, including changes in attitudes, beliefs, feelings and behavior that result from the comments, actions, or even the mere presence of others." [55].

Several studies have been made for analyzing traits and behaviors to identify the key characteristics of influential along three lines [75]: i) who one is the individual characteristics of opinion leaders (i.e.personality traits, charisma or demographic and socioeconomic backgrounds); ii) what one knows , the characteristics pertaining to individuals' competence, such as their knowledge, expertise, or ability to provide information or guidance on particular issues; and iii) whom one knows , the characteristics related to an individual's structural position in a network.

Traditional communication theories state that a minority of users, called *influentials*, excel in persuading others. One of the most challenge problems in literature is the *influence maximization* problem. This problem has applications in viral marketing, where a company may wish to spread a new product via the most influential individuals in popular social networks. Richardson and Domingos [41] deal with a first problem of influence maximization related to Viral Marketing application. In particular, they examine a particular market as a social network composed by different interconnecting entities and model it as a Random Markov field in order to identify a subset of users to convince to adopt a new technology for maximizing the adoption of this technology. In this problem, a social network is modeled by a graph $G(V, E, \omega)$,where

V is the set of OSN users, $E \subseteq V \times V$ the set of direct relationships(i.e., friendship, following/follower etc.) between users and $\omega : V \times V \rightarrow [0,1]$ is a function that assigns to each edge a number between 0 and 1.

The objective in influence maximization (IM) is to maximize the spread of information in a network through activation of an initial set of k seed nodes. To this end, we first define the notions of seed and active nodes.

Definition 2.4.1 (Seed node)**.** *Let $G = (V, E)$ be an Online Social Network, a* seed node *is a node $v \in V$ that acts as the source of information diffusion in the OSN. The set of seed nodes is denoted by S.*

Definition 2.4.2 (Active node)**.** *Let $G = (V, E)$ be an Online Social Network, a node $v \in V$ is a* active node *if either (1) It is a seed node ($v \in S$) or (2) It receives information under the dynamics of an information diffusion model I, from a previously active node $u \in V_a$. Once activated, the node v is added to the set of active nodes V_a.*

Thus, an *influence analysis* problem can be faced using two steps.

In the first one, a *diffusion model* describes the dynamics of how information spread in OSN; this phenomenon is usually modeled by a stochastic process where the activation of each node is based on its neighbours state. In the second step, a *maximization algorithm* is exploited to identify the set of nodes such that their activations maximize the diffusion or the propagation of influence.

Different models have been proposed in literature, that can be classified into two categories: *Progressive* and *Non-progressive* models. The difference existing between the two models is due to the activated nodes in the *non progressive* models can be switched their state indefinitely.

Based on the spread models, the influence maximization problem has been proposed in literature. The selection of the most influence nodes is an optimization problem that has been proven by Kempe et al.[76]

to be NP-Hard. The influence function $\sigma(S)$ maps subsets of elements of a finite set to a non-negative number denoting the expected size of the activated set if S is targeted for initial activation. The final goal is to find a k-element seed set S that maximizes $\sigma(S)$. In particular, chosen S as the initial active seed-set, Kempe et al. defined its influence $\sigma(S)$ as the total number of activated nodes in the network at the end of the diffusion process. To address this complexity, they exploit a result of Nemhauser, Wolsey, and Fisher [33][100] to show that a greedy hill-climbing algorithm approximates the optimum to within a factor of $(1-\frac{1}{e})$ (where e is the base of the natural logarithm): start with the empty set, and repeatedly add an element that gives the maximum marginal gain.

Theorem 2.4.1. *For a non-negative, monotone submodular function f, let S be a set of size k obtained by selecting elements one at a time, each time choosing an element that provides the largest marginal increase in the function value. Let S^* be a set that maximizes the value of f over all k-element sets. Then $\sigma(S) \geq (1-\frac{1}{e}) \cdot \sigma(S^*)$; in other words, S provides a $(1-\frac{1}{e})$-approximation.*

It is possible to define σ as submodular if it satisfies a natural diminishing returns property: the marginal gain from adding an element to a set S is at least as high as the marginal gain from adding the same element to a superset of S. Formally, a submodular function satisfies:

$$\sigma(S \cup \{v\}) - \sigma(S) \geq \sigma(T \cup \{v\}) - \sigma(T)$$

for all elements v and all pairs of sets $S \subseteq T$.

Submodular functions have a number of very nice tractability properties; the one that is relevant to us here is the following. Suppose we have a function σ that is submodular, takes only nonnegative values, and is monotone in the sense that adding an element to a set cannot

cause f to decrease: $\sigma(S \cup \{v\}) \geq \sigma(S)$ for all elements v and sets S. We wish to find a k-element set S for which $\sigma(S)$ is maximized. This is a performance guarantee slightly better than 63%.

The algorithm 1 that achieves this performance guarantee is a natural greedy hill-climbing strategy. Starting from the empty set, it repeatedly adds to A the node x maximizing the marginal gain $\sigma(A \cup \{x\}) - \sigma(A)$. However, it is possible to obtain arbitrarily good approximations to $\sigma(\cdot)$ in polynomial time, simply by simulating the random choices and diffusion process sufficiently many times.

Algorithm 1 Greedy Approximation Algorithm

```
1: procedure GREEDY APPROXIMATION ALGORITHM(G,k)
2:     S ← 0
3:     while do(|S| ≤ k)
4:         j ← argmax_{v∈V}(σ(S∪v) − σ(S))
5:         S ← S∪{j}
6:     end while
7:     return S
8: end procedure
```

Moreover the authors exploit the Montecarlo simulation to provide a better accurate approximation of influence spread, which turns out to be inefficient for large networks. This weakness led to develop different families of methods to estimate the influence spread: *simulation based*, *heuristic based* and *sketch based*. The first family [27, 83, 103] has the aim to repeat the simulation method to improve the accuracy of influence spread process. The second and the third families try to overcome the inefficiency of Montecarlo simulation using two different approaches: *heuristic based* methods[70, 146] exploit communities or linear systems for restricting the influence spread while *sketch based* techniques[22, 133] build a family of vertices sets exploiting the reverse simulation.

2.4.2 Community Detection

The continuous increase of the number of people that sign up in On-line Social Networks implicates that the interaction among users are complex to manage. Moreover, the growth of OSNs has increased the amount of heterogeneity information produced or consumed by users; in fact, for instance, Facebook has 1,870 milion active users producing 510.000 comments and 136,000 each 60 seconds, Youtube have 1,300,000,000 of people that upload 300 hours of video every minute. These facets led complex to understand how the interactions established among users can exert influence about user preferences.

Thus, the needs to detect users' communities based on his interests and his social connection became an important challenge in literature and can support several applications; in fact, the diffusion of a new idea or technologies can be maximized for identifying of people group interested about a given topic, the recommendation suggestion can be improved taking in account also how social ties can be influenced by users' behaviors in the same communities.

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real systems is community structure, or clustering, i. e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, can be considered as fairly independent compartments of a graph, playing a similar role like, e. g., the tissues or the organs in the human body. Detecting communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs. This problem is very hard and not yet satisfactorily solved, despite the huge effort of a large interdisciplinary community of scientists working on it over the past few years. In graph theory community detection is a

fundamental task to analyze and understand the structure of complex network. Revealing the latent community structure, which is crucial to understanding the features of networks, is an important problem in network and graph analysis. During the last decade, many approaches have been proposed to solve this challenging problem in diverse ways. The term "community" has been extensively used in the literature in different contexts and with different connotations. Since there is no single definition of a community which is universally accepted, many thoughts emerged which in turn resulted in different definitions of community structure.

A first definition is provided by Gulbahce and Lehmann [62] that define a community as "a densely connected subset of nodes that is only sparsely linked to the remaining network". Porter [108] provides a definition based on the fields of sociology and anthropology, in which a community is "cohesive groups of nodes that are connected more densely to each other than to the nodes in other communities". Moreover, Fortunato [48] defines communities as "groups of vertices that probably share common properties and/or play similar roles within the graph. The most commonly used definition is that of Yang [154]: "a community is a group of network nodes, within which the links connecting nodes are dense but between which they are sparse". Papadopoulos [105] ultimately defines communities as: "groups of vertices that are more densely connected to each other than to the rest of the network".

Fortunato [48] identifies three levels to define a community: (i) local definitions, "parts of the graph with few ties to the rest of the system"; (ii) global definitions, a global criterion based on the chosen algorithm and associated with the graph is used to compute communities ; and (iii) definition based on vertex similarity, communities are considered as groups of vertices similar to one another.

However, the definition of a community is not clear because it is complex to define the concept of good community. Several possible communities can be computed in a social network $G(V, E)$ whose enumeration is a NP-Complete problem [54]. The aim of several community detection algorithm is to optimize a goodness metric that represents the quality of the communities detected from the network.

In ONSs it is easy to note that each user can belong to multiple communities, whose number is potentially unlimited; for instance a user is connected to different groups of people that represent its family, collegues, friends and so on.

Thus the problem of community. In a MSN it is possible to define two types of community detection problem:

1. without overlapping: corresponds to partition a vertices set into k pairwise disjoint clusters $C_1, \ldots, C_k$ such that $C_1 \cup \ldots \cup C_k = V$, where V is the vertices set.

2. with overlapping: corresponds to identify overlapping clusters, called cover C = c1, c2,..., ck [80], whose union is not necessarily equal to the entire vertex set V, that is $C_1 \cup \ldots \cup C_k \subseteq V$.

A first classification of community detection approaches is based on the type of model used to represent the OSN. The first two approaches model the OSN respectively as a graph, the most widespread, and as a hypergraph by exploiting the well-known properties and algorithms on such models. The last family exploits the lattices concepts and algorithms based on lattices and hypergraph structures for analyzing OSNs.

Usually a social network is defined as a graph G=(V,E), where V is a set of nodes and E is a set of edges

It is possible to define the following two types of community:

1. A strong community is a subgraph in which each vertex has more likehood to be connected with the vertices of the same community compared to a node in another community.

2. A weak community is a subgraph in which the average edge probability between each vertex in the same community exceeds the average edge probability between the vertex with the vertices of any other group.

Thus, the two definitions are based on different granularity levels: strong communities is defined for each pair of nodes while the weak community is defined for averages over groups.

As shown in [26], the community detection analysis is composed by two phases: detection of meaningful community structure from a network, and second, evaluation of the appropriateness of the detected community structure. In particular, the second phase raises the importance to define several metrics that are the fundamentals of different community detection definitions.

For this reason, several measure can be defined to evaluate the goodness of a community:

One of the most popular metrics is Modularity, that measures the number of edges that lie within a cluster compared to the expected number of edges of a null graph. In other words, identifying of community structures is related to finding communities that maximize modularity values because an high value of modularity corresponds to better community structure. Formally, the modularity measure of a specific community in undirect graph is defined as follows:

$$Q_u = \frac{1}{2m} \sum_{i,j} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j) \tag{2.1}$$

where A is the adjacency matrix, c_i $(\forall i \in V)$ the community membership of node i and $\delta(c_i, c_j)$

Another important metrics to evaluate the goodness of community is the Normalized Mutual Information (NMI), that evaluates information-theoretic concepts based on the concept that it is possible to infer a cluster assignment with respect to another on with small additional information if the two examined clusters are close. Formally, it is possible to define NMI as follows:

$$NMI(A,B) = \frac{-2\sum_{i}^{C_a}\sum_{j}^{C_b} N_{ij}\log(\frac{N_{ij}N}{N_i N_j})}{\sum_{i}^{C_a} N_{i:}\log(\frac{N_i}{N}) + \sum_{j}^{C_b} N_{:j}\log(\frac{N_i}{N})} \tag{2.2}$$

where $|C_A|$, $|C_B|$ represent the number of ground truth clusters and the number of produced clusters respectively. The element N_{ij} corresponds to the number of nodes in real cluster i that appear in the produced cluster j, while N_i is the sum over row i and N;j the sum over column j of the confusion matrix N. In the case where the produced results are identical with the ground truth, the NMI(A, B) measure takes its maximum value one, while in the case where the two clusterings totally disagree, the NMI(A, B) score is zero.

Moreover community detection approaches can be classified in two categories: global, that require the knowledge of entire network, and local algorithms, that expand an initial seed set of nodes into possibly overlapping communities by examining only a small part of the network.

The proposed approaches for community detection can be classified in the following five categories:

1. Cohesive subgraph discovery
2. Vertex clustering
3. Community quality optimization
4. Divisive
5. Model based

Finally, in the following table[1][2] the proposed community detection algorithms are summarized, as shown in [95].

[1]In the "Type" column, L and G denote local and global, and H and NH denote hierarchical and non-hierarchical, respectively. The LG algorithm can find hierarchical communities if the node-based algorithm is hierarchical.

[2]In the "Complexity" column, n denotes the number of nodes, m denotes the number of edges, K is the maximum node degree, t is the number of algorithm iterations selected, α is the power-law exponent, $vol(C)$ is the sum of the degree of all the nodes in a community C, and C is the set of all the identified communities.

	Algorithm	Type	Description	Complexity
Non-overlapping	Blondel[20]	G	Fast modularity maximization (Louvain) is a greedy approach to modularity maximization and unfolds a hierarchical community structure.	O(m)
	Infomap[119] InfoH[120]	G	Maps of random walks finds communities based on the compression of the description length of the average path of a random walker over the network. Multilevel compression of random walks is the hierarchical version of Infomap which minimizes a hierarchical map equation to find the shortest multilevel description length.	O(m)
	RN[65]	G	Potts model community detection minimizes the Hamiltonian of a local objective function (the absolute Potts model).	O(m)
	MCL[143]	G	Markov Clustering is based on the probability of random walks remaining for a long time in a dense community before moving to another community.	$O(nK^2)$
Overlapping	LC[2]	G	Link Community detection uses the similarity of the edges to identify hierarchical communities of edges rather than communities of nodes.	$O(nK^2)$
	LG[46]	G	Line Graph and graph partitioning runs a non-overlapping node-based algorithm on a linegraph induced from the original graph to identify overlapping link-based communities.	O(nm^2)
	SLPA[153]	G	Speaker listener Label Propagation is an extension to the label propagation algorithm where nodes adopt multiple labels based on the majority labels in their neighborhood.	$O(tm)$
	OSLOM[80]	L	Order Statistics Local Optimization Method identifies significant communities with respect to a Null model similar to modularity.	$O(n^2)$
	DEMON[35]	L	Democratic Estimate of the Modular Organization of a Network is a local algorithm which uses the label propagation algorithm to find communities in the egonet of each node and then merges them into larger communities.	$O(nK^{3-\alpha})$
	PPR[156]	L	Personalized PageRank-based, is a local algorithm which uses the PageRank-Nibble algorithm [10] to approximate a personalized PageRank vector from a given seed node and then uses the method in [14] to create the communities based on a scoring function.	$O(\sum c \in CVol(c))$

Table 2.1: Community Detection Approaches

3

Multimedia Social Network Model

Nowadays, Multimedia Social Networks (MSN) have attracted a large number of users, providing them new features that allow users to interact and share multimedia contents with their colleagues, families and friends.

MSNs provide new interaction ways that integrate sociology multimedia and communication technologies. In a MSN, thus, several types of relationships can be made between users, multimedia objects and users and multimedia objects; in particular, users can make friendship, create groups while multimedia relationships based on similarity high-level and low-level features can be instantiated between multimedia contents. Finally, users interact with multimedia contents in several ways: share, comments or rating for a multimedia contents or choose some of them as favorite.

The richness and quantity of entities and relations produced by MSNs provides many opportunities for *Social Big Data analysis*. In particular, Social Big Data can be defined as the set of of processes and methods designed to analyze social media data sources, that can be characterized by their heterogeneous information, size and the on-line or streamed generation of information, with the aim to provide sensitive and relevant knowledge to any user or company.

3.1 Modeling MSNs using hypergraph

The growth of MSNs led to produce heterogeneous information with the related metadata on the network. In this section a novel data model is provided for describing MSNs[8]. In particular, the proposed model is composed by two types of vertices: users and multimedia objects.

A multimedia object entity represents the item used by users as means to spread its idea or interact with other users; for instances, in Facebook or Twitter, a multimedia object can be seen as a post, image or video shared by a user while in business oriented networks it corresponds to a review or a tip about a given business object. For this reason, we define four types of multimedia objects based on given examined multimedia contents: *text*, *video*, *audio* and *images*. In the most diffused online large-scale social networks, as Twitter, Google$^+$ and Facebook, multimedia objects are represented by different types of heterogeneous content such as tweets, posts, but also videos, photos and so on. In other kinds of social networks (e.g. Instagram, Flickr, Youtube, Last.fm, etc.) objects are essentially multimedia data (i.e. images, video and audio contents). Moreover, in some interest-based social networks (e.g. Yelp, Imdb, etc.) or location-based social networks (e.g. TripAdvisor, Foursquare) multimedia objects correspond to specific items that are usually rated for recommendation purposes (e.g restaurants in TripAdvisor, movies in IMDB, business objects in Yelp, places of interest in Foursquare, etc.). Multimedia objects can be obviously described using *metadata* and different annotation schema. In addition, in our model also multimedia data low-level *features* can be properly used. Eventually, it is possible to derive another type of multimedia object, called *Annotation Assets* from the analysis of textual annotations, mainly tags but also keywords, comments, reviews related to items published on the MSN. In particular, they represent the most significant terms or named entities whose definition can be retrieved from dictionaries,

ontologies and so on - of one or more domains, exploited by users to describe multimedia objects.

User is a collection of person and organizations constituting one or more social communities. Several information concerning their profile, interests, preferences, etc. can be opportunely taken into account. Each user can perform several actions on the network: create or interact with different contents in the network, browse other user profiles or images. This entity can be composed by several information describing its biography information, personal preferences and so on; for instance in the well-known OSNs (e.g. Twitter, Instagram, Facebook etc.) a user is characterized by a username, typically associated to personal name, age and its preferences as well as reading books, listening to music, watching movies while in business oriented social network (i.e. Yelp, Trip advisor) it is also associated a trustworthiness value defined by several measures, such as mean votes about of its reviews number and type of received feedbacks etc.

The aim is to provide a general model allowing to represent in an effective way any kind of relationships in any type of MSNs. To properly manage the heterogeneity of well-known MSNs an hypergraph data structure is proposed. The proposed model leverages this structure because it allows to represent complex relationships that it is possible to establish in a Multimedia Social Network (MSN). In fact, several types of relationships can be established among the entities involved in Multimedia Social Network model. For example, a user can publish an image, two friends can comment the same post, a user can tag another user in a photo, a user can listen a song, a user can share some videos within a group, a user can write a review about a restaurant and so on, just to make several examples. To properly manage the heterogeneity of well-known MSNs an hypergraph data structure is proposed. The proposed model leverages this structure because it allows to represent

complex relationships that it is possible to establish in a Multimedia Social Network (MSN).

Thus, due to the variety and intrinsic complexity of these relationships, we decided to leverage the *hypergraph* formalism to model an OSN from a conceptual/logic point of view. In particular, our model relies on several concepts: i) an OSN is seen as particular a weighted and undirected hypergraph; ii) by means of the discussed relationships proper social paths (i.e. hyperpaths) permit to "connect" the entities of our model. In the following, we provide all the basic definitions that characterize our model.

Definition 3.1.1 (MSN). *Let $HV = U \cup O$, be a finite set of* vertices*,* U *being the set of OSN users and* O *the set of objects. Let HE be a set of* hyperedges *with a finite set of indexes I. Let $\omega : HE \to [0,1]$ a weight function. A* Multimedia Social Network *is the triple $\langle HV; HE; \omega \rangle$.*

In particular, we assume that each hyperedge $he_i \in HE$ is in turn defined by a ordered pair $he_i = (he_i^+ = (HV_{he_i}^+, i); he_i^- = (i, HV_{he_i}^-))$. The element he_i^+ is called the *tail* of the hyperarc he_i whereas he_i^- is its *head*, $HV_{he_i}^+ \subseteq HV$ being the set of vertices of he_i^+, $HV_{he_i}^- \subseteq HV$ the set of vertices of he_i^- and $HV_{he_i} = HV_{he_i}^+ \cup HV_{he_i}^-$ the subset of vertices constituting the whole hyperedge.

An hypergraph related to an HSN can be also denoted by an *incidence matrix* H with entries as:

$$h(v, e_i) = \begin{cases} 1, & \text{if } v \in V_{e_i} \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

Actually, vertices and hyperedges are *objects* with a set of properties that allow to model several social networks, at the same time supporting different applications. For this reason, a '' dot notation" is used to identify the attributes of a given vertex or hyperedge: as an example,

he_i.id, he_i.name, he_i.time, he_i.source and he_i.type represent the id, name, timestamp, source (social network) and type of the hyperedge he_i, respectively. Moreover, it is possible to note that several attributes are common to any kind of vertices/hyperdges, in turn, other ones are typical of specific vertices/hyperdges. Eventually, the weight function can be used to define the "confidence" of a given relationship.

3.1.1 Relationships

It is possible to classify the relationships in a MSN using the following categories:

- **User to User** relationships, describing user actions towards other users;
- **Similarity** relationships, describing a relatedness between two objects or users;
- **User to Object** relationships, describing user actions on objects, eventually involving some topics or other users.

In the following, it is provided the definition for each type of relationship.

Definition 3.1.2 (User to User relationship)**.** *Let* $\widehat{U} \subseteq U$ *be a subset of users, we define a* user to user relationship *each hyperedge* he_i *having the following properties:*

1. $HV^{+}_{he_i} = u_k$ *such that* $u_k \in \widehat{U}$,
2. $HV^{-}_{he_i} \subseteq \widehat{U} - u_k$.

Examples of "user to user" relationships are properly represented by *friendship*, *following* or *membership* of some famouse online social

networks. For this kind of relationships, $\omega(he_i) \in [0,1]$ depending on the particular application. In the opposite, a general strategy can assign the value 1 to each user to user relationship.

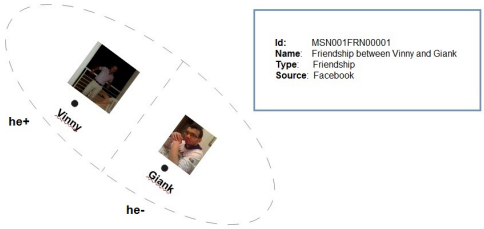


Figure 3.1: Friendship relationship.

Example 3.1.1 (Friendship relationship)**.** *Let* Vinny *and* Giank *be two users in a MSN, Figure 3.1 shows how it is easily to represent the* Friendship relationship *between these two users; in particular, Vinny represents the user that sends the friendship request while Giank is the user that accepts the request.*

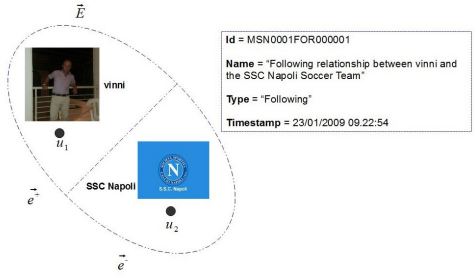


Figure 3.2: Following relationship.

Example 3.1.2 (Following relationship)**.** *Let Vinny and Giank be two different users in a MSN, an example of* following relationship *is shown in figure 3.2 in which in particular, Vinny is the user who desires to follow activities of another user, Giank, that is the following user.*

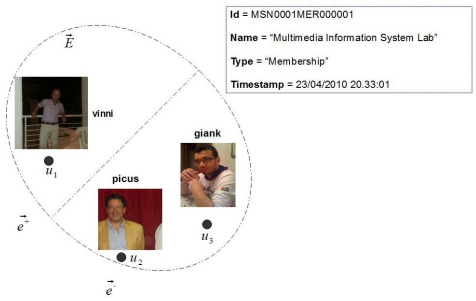


Figure 3.3: Membership relationship.

Example 3.1.3 (Membership relationship)**.** *Let Vinny, Giank and Picus be three different users in a MSN, an example of* membership relationship *is shown in figure 3.3 in which Vinny is the group founder while the other users, included in the* $HV_{e_i}^{+}$*, are the members.*

This function can be applied to a pair of vertices of the same type, and in particular it is possible to compute a similarity value: between two users by considering different types of features (interests, profile information, preferences, etc.), between two multimedia objects using the well-know (high and low level) features and metrics proposed in the literature, between two annotation assets exploiting the related concepts and the well-know metrics on vocabularies or ontologies.

Moreover, it is possible to compute a similarity value: between two users (by considering different types of features (interests, profile information, preferences, etc.); between two objects (using the well-known high and low level features and metrics proposed in the literature)

Definition 3.1.3 (Similarity relationship)**.** *Let* $v_k, v_j \in V$ $(k \neq j)$ *two vertices of the same type of an OSN, we define* similarity relationship *each hyperedge* he_i *with* $HV_{he_i}^{+} = v_k$ *and* $HV_{he_i}^{-} = v_j$*. The weight function for this relationship returns similarity value between the two vertices.*

In the proposed model, a similarity hyperedge is effectively generated if $\omega(\vec{he}_i) \geq \eta$, η being a given threshold (see Figure 3.4).

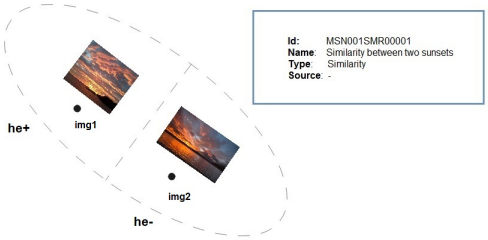


Figure 3.4: Multimedia similarity relationship.

Finally, the last type of relationships is the *User to Object relationship.*

Definition 3.1.4 (User to Object relationship)**.** *Let $\widehat{U} \subseteq U$ a set of users and $\widehat{O} \subseteq O$ a set of objects in a OSN, we define* user to object relationship *each hyperedge he_i with the following properties:*

1. $HV^{+}_{he_i} = u_k$ *such that* $u_k \in \widehat{U}$,
2. $HV^{-}_{he_i} \subseteq \widehat{O} \cup \widehat{U}$.

Examples of "user to object" relationships are the following: *publishing*, *reaction*, *annotation*, *review*, *comment* (in the last three cases the set $HV^{-}_{e_i}$ can also contains one or more topics) or *user tagging* (involving also one ore more users) activities. For this kind of relationships, the $\omega(he_i)$ is set to a value in [0,1] that is function of the specific relationship and depends on the particular supported application. In the opposite, a general strategy can assign the value 1 to each user to object relationship.

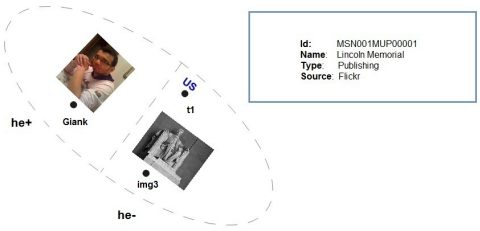


Figure 3.5: Multimedia tagging relationship.

Example 3.1.4 (Publishing relationship)**.** *Let Giank and* Im_1 *be respectively a user and a multimedia object, an example of* Publishing relationship *is shown in figure 3.5, in which Giank is the user who desires to publish the multimedia object* Im_1*. This relationship takes place in a MSN when a user publishes several multimedia object to share it with other users or to remind a particular event.*

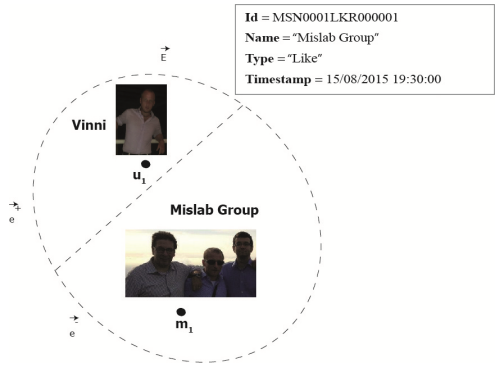


Figure 3.6: Like relationship.

Example 3.1.5 (Like relationship)**.** *Let Vinni and* Im_1 *be respectively a user and a multimedia object of a MSN, an example of* Like relationship *is shown in Figure 3.6; in particular in the set* $HV_{e_i}^{+}$ *there is the user (Vinni) who likes the* m_j *multimedia object in the* $\vec{e}_k.\mathsf{time}$ *instant. This kind of relationship is typical of the Facebook OSN.*

3.2 MSN Operations

In this section we analyze the operations that are possible to perform on the proposed MSN model. Firstly we define the basic operations and successively we analyze how it is possible to combine them for developing some algorithms.

In our model we consider the dynamic behaviors of OSN using the following basic operations.

Definition 3.2.1 (Vertex Insert)**.** *Let $MSN = (HV, HE)$ be a Multimedia Social Network and hv a new node, a* vertex insert *corresponds to carry out the following operation:*

$$HV = HV \cup hv$$

A vertex insert operation corresponds to add a new element in the set of entity involved in a MSN. In particular, several properties can be defined for each new node depend on which entity the occurrence belongs to.

Definition 3.2.2 (Edge Insert)**.** *Let $MSN = (HV, HE)$ be a Multimedia Social Network and $he \subseteq HE$ an hyperedge, an* hyperedge insert *corresponds to perform following operation:*

$$HE = HE \cup he$$

The insert of a new hyperedge in the MSN corresponds to add either a new action performing from a user belonging to the HV_{he}^{+} set with respect to other entities or a relationship based on similarity value computed between two occurrences of the same entity.

Thus, the insertion of a new object (user or multimedia object) into a MSN corresponds to the execution of the two operations described above.

Definition 3.2.3 (Insert Object)**.** *Let $MSN = (HV, HE)$ be a Multimedia Social Network, hv and $HE_v \subseteq HE$ be respectively a new node and a sequence of actions made by it, an* insert object *corresponds to perform following operation:*

$$HV = HV \cup hv$$

$$HE = HE \cup HE_v$$

In turn, the delete operation of a vertex led to update the MSN by a two phase procedure: firstly it removes each hyperedge in which the analyzed node is involved and successively the node itself is removed.

Definition 3.2.4 (Vertex Delete)**.** *Let $MSN = (HV, HE)$ be a multimedia social network and $hv \in HV$ a vertex, a* Vertex deletion *corresponds to the update operation on the MSN in following way:*

$$MSN = (HV', HE') \rightarrow$$

$$\begin{cases} HV` \setminus hv \\ HE' = \{he \in HE : hv \notin HV_{he}^{+} \wedge hv \notin HV_{he}^{-}\} \end{cases}$$

The hyperedge deletion allows to delete an action made by a user or a similarity relationship established between two entities because their preferences are changed or other content analysis are made on them.

Definition 3.2.5 (HyperEdge Delete)**.** *Let $MSN = (HV, HE)$ be a Multimedia Social Netework and $he \in HE$ an hyperedge, an* hyperedge deletion *corresponds to the following operation:*

$$HE = HV \setminus he$$

Once defined the basic operations, we develop some algorithms exploiting the previous definition.

Definition 3.2.6 (Subgraph)**.** *Let $MSN = (HV, HE)$ be a Multimedia Social Network and Θ a set of conditions defined over the attributes of vertices and hyperedges, a* subgraph *corresponds to a sub-hypergraph $MSN^{'} = (HV`, HE`)$ defined as follow:*

$$MSN` = (HV', HE') \rightarrow$$

$$\begin{cases} HV^{`} \subseteq HV = \{hv \in HV^{`} : hv \vDash \theta\} \\ e_j \in HE^{`} : e_j \subseteq HV^{`} \wedge e_j \vDash \theta \end{cases}$$

In the modern On-line Social Networks, the interactions among users occur by several ways, that can involve also the multimedia contents shared on the network. To describe how two users interact between them, we introduce the *Social Path* concept.

Definition 3.2.7 (Social Path)**.** *A* social path *between two vertices* v_{s_1} *and* v_{s_k} *of an OSN is a sequence of distinct vertices and hyperedges* $sp(v_{s_1}, v_{s_k}) = v_{s_1}, he_{s_1}, v_{s_2}, ..., he_{s_{k-1}}, v_{s_k}$ *such that* $\{v_{s_i}, v_{s_{i+1}}\} \subseteq HV_{he_{s_i}}$ *for* $1 \leq i \leq k-1$*. The* length γ *of the hyperpath is* $\alpha \cdot \sum_{i=1}^{k-1} \frac{1}{\omega(he_{s_i})}$*,* α *being a normalizing factor. We say that a social path* contains *a vertex* v_h *if* $\exists he_{s_i} : v_h \in he_{s_i}$*.*

Social paths between two nodes leverage the different kinds of relationships: a given path can "directly" connect two users because they are "friends" or members of the same group, or "indirectly", as they have commented the same picture.

To analyze how the actions made on MSN by users are correlated, it is important to define the concept of relevant path. A relevant path corresponds to admissible sequences of hyperedges on which a given user can spread his influence respect to another one. Note that, with abuse of notation, $v_i \underset{\tau}{\rightarrow} v_j$ iff $\exists u_i \in V_i \wedge u_k \in V_j : u_i \underset{\tau}{\rightarrow} u_j$.

Definition 3.2.8 (τ-Relevant path)**.** *A relevant path between two vertices* v_{s_1} *and* v_{s_k} *of a MSN is a sequence of distinct vertices and hyperedges* $v_{s_1}, e_{s_1}, v_{s_2}, ..., e_{s_{k-1}}, u_{s_k}$ *such that* $\{v_{s_i}, v_{v_{i+1}}\} \subseteq V_{e_{s_i}}$ *for* $2 \leq i \leq k-1 \wedge$ $v_{s_i} \underset{\tau}{\rightarrow} v_{s_{i+1}}$*.*

3.2.1 Centrality Measures

One of the fundamentals in Social Network Analysis is to compute the *centrality* measures for the user nodes of a social graph. As well known,

the centrality represents the "importance" of a given user within the related community and can be easily exploited for several applications: influence analysis, expert finding, communities detection, recommendation, etc., just to make some examples. In the literature [71, 1, 38], there exist a lot of measures to determine the centrality of a node in a social graph.

In [7], the most diffused ones (in the case of undirected graphs) are defined on the proposed hypergraph-based MSN model and a new centrality measure based on the concept of "neighborhood" among users is proposed.

Definition 3.2.9 (Degree Centrality)**.** *Let $v_k \in V$ a vertex of an HSN and H the related incidence matrix, we define the* degree centrality *of v_k as:*

$$dc(v_k) = \sum_{e_i \in E} h(v_k, e_i) \tag{3.2}$$

The following example is shown to better explain the meaning of this measure.

Example 3.2.1. *Degree Centrality Let us consider the subnet described in Figure 3.7*[1]

The degree centrality of a node depends on the number of social relationships (i.e. hyperedges) in which the node is involved; for instance, the user Giank *has degree equals to* 5.0 *because it is involved in five relationships.*

Definition 3.2.10 (Closeness Centrality)**.** *Let $v_k \in V$ a vertex of an HSN, we define the* closeness centrality *of v_k as:*

$$cc(v_k) = \frac{1}{\sum_{v_j \in V} d_{min}(v_k, v_j)} \tag{3.3}$$

[1]The most important topics within the reviews are considered.

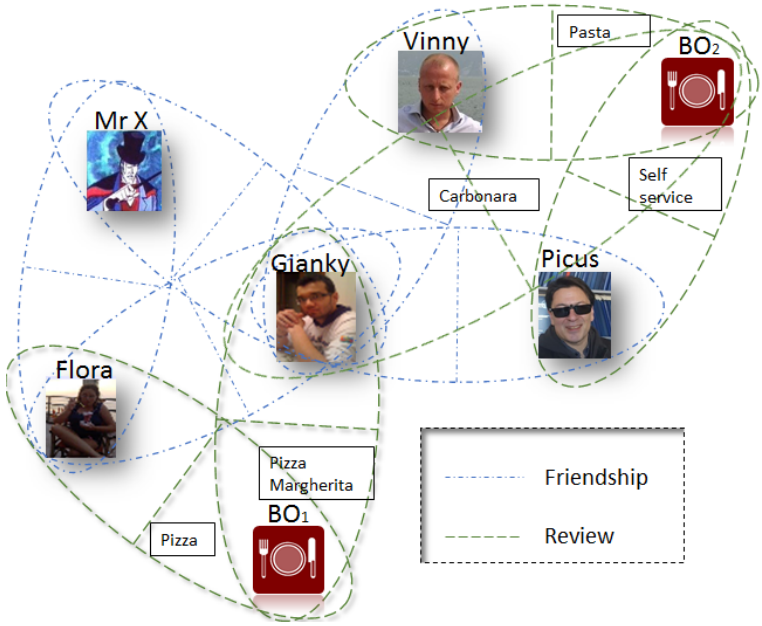


Figure 3.7: Example of HSN

The meaning of the *Closeness Centrality* is better explained through the following example.

Example 3.2.2. *Closeness Centrality Let us consider the subnet described in Figure 3.7. The closeness centrality of a node depends on the sum of distances respect to all the other nodes; in particular, the nodes from which it is more simple to reach the other ones are those more important. Thus, in Figure 3.7 the users* Vinni *and* Picus *have the same closeness value (equal to* 0.14*) because they have the same distances from the other nodes in the network.*

Definition 3.2.11 (Betweenness Centrality)**.** *Let $v_k \in V$ a vertex of an HSN, we define the* betweenness centrality *of v_k as:*

$$bc(v_k) = \frac{\sum_{v_j \neq v_z \in V} \sigma_{v_j v_z}(v_k)}{\sigma_{v_j v_z}} \tag{3.4}$$

$\sigma_{v_j v_z}(v_k)$ being the number of shortest hyperpaths connecting v_j, v_z and passing through v_k, while $\sigma_{v_j v_z}$ is the total number of shortest hyperpaths between v_j, v_z,

The following example is shown to better explain the meaning of this measure.

Example 3.2.3. *Betweenness Centrality Let us consider the subnet described in Figure 3.7. The betweenness centrality measures the number of times that a node is present in the shortest hyperpaths between each pair of distinct vertices of a hypergraph. It is easy to note that the user* Giank *has the highest value of betweenness since all shortest paths to reach other users involve it.*

In addition to the discussed measures, we have introduced a novel centrality measures that exploit the concept of λ-Nearest Neighbors Set.

Definition 3.2.12 (Neighborhood Centrality)**.** *Let $v_k \in V$ a vertex of an HSN and λ a given threshold, we define the* neighborhood centrality *of v_k as:*

$$nc(v_k) = \frac{\left|NN_{v_k}^{\lambda} \cap V\right|}{|V| - 1} \tag{3.5}$$

NN_k^{λ} being the λ-Nearest Users Set of v_k.

The neighborhood centrality of a given node can be measured by the number of nodes that are "reachable" within a certain number of steps using social paths.

Except for the degree centrality, the other introduced measures can be computed locally with respect to a community of users ($\widehat{U} \subseteq U \subseteq V$) and considering only vertices of user type for the end-to-end nodes of hyperpaths. In this manner, centrality is referred to user importance within the related community. We define *user centrality* such kind of measure.

In addition, in order to give more importance to user-to-content relationships during the computation of distances for the user neighborhood centrality, it is possible to apply a *penalty* if the considered hyperpaths contain some users; in this way, all the distances can be computed as $\tilde{d}(v_k, v_j) = d(v_k, v_j) + \beta \cdot N$, N being the number of user vertices in the hyperpath between v_i and v_j and β a scaling factor[2].

Eventually, it is possible to compute a *topic-sensitive centrality* for the closeness, betweenness and neighborhood measures considering in the distances' computation only hyperpaths that contain a given topic node.

Just as an example, the *topic-sensitive user neighborhood centrality* for a given user community is:

[2]Such strategy is necessary to penalize *lurkers*, i.e. users of an HN that do not directly interact with content but through user to user relationships.

	Degree	Closeness	Betweenness	Neighborhood ($\lambda = 2, \beta = 1.5$)
Vinny	2.00	0.14	0	0.25
Picus	2.00	0.14	0	0.25
Giank	5.00	0.25	0.6	0.75
Mister X	2.00	0.16	0	0.0
Flora	3.00	0.16	0	0.25

Table 3.1: Example of user centrality measures

$$nc(u_k|\widehat{U}, t_z) = \frac{\left|NNU^{\lambda}_{u_k t_z} \cap \widehat{U}\right|}{\left|\widehat{U}\right| - 1} \tag{3.6}$$

$\widehat{U}$ being a user community, u_k a single user and t_z a given topic.

Example 3.2.4. *Neighbor centrality Let us consider the subnet described in Figure 3.7. Considering the proposed Neighborhood measure, it is easy to note that the node* Mister X *has a minimal neighborhood centrality linked to a node having maximal neighborhood centrality.*

Table 3.1 reports the user centrality measures' values related to the MSN representation.

4 Algorithms

In the last years, the emergence and progress of Multimedia Social Networks (such as Facebook, Twitter, Flickr and so on) allowed users to interact, engage and share multimedia contents with each other by several actions such as reply, comment, subscribe and connect.

In particular, the exponential growth of multimedia contents, produced by the ubiquitous presence of different devices such as phones, digital cameras, and camcorders, in MSNs led to define new challenges about their storage and management due to their high computational complexity. For this reason, novel methodologies and algorithms have been proposed for Influence Maximization and Community detection problems in the following sections.

4.1 Influence Analysis

The growing popularity of MSNs and, in particular, their huge amount of data and social interactions, lay the foundations for analyzing several sociological phenomena that can be used for a great number of applications. From this perspective, MSNs are social structures composed by a set of actors (individuals or organizations), sets of dyadic ties, and other

social interactions, often instantiated through the shared information, among actors themselves. In this case, mathematical models can be adopted to study the social network structure, the related generated content and its temporal evolution.

Despite the great amount of research done in the Multimedia Social Networks (MSNs) field, only few works have investigated the use of multimedia data in such realm. Instead, a novel data model, shown in the previous chapter, that takes into account the intrinsic characteristics of multimedia may be of great help in managing Multimedia MSNs for providing more effective algorithms in a variety of applications, especially for Influence Analysis aims.

In particular, the proposed strategy to deal with this problem is composed by two phases: firstly an *Influence Model* has been introduced and successively different algorithms have been developed.

4.1.1 Influence Model

In the proposed model, the basic assumption is the existence of a finite set of *Actions* (A) representing the "interactions" among a finite set of *Users* (U) and a finite set of *Objects* (O) in several On-line Social Networks: they can be properly captured during user browsing sessions exploiting the log information.

Definition 4.1.1 (Log tuple)**.** *A* log tuple *is defined by* $l = (a, u, o, \lambda_1, \cdots, \lambda_k)$, $a \in A$, $u \in U$, $o \in O$ *and* $\lambda_1, \cdots, \lambda_k$ *being particular attributes (e.g., timestamp, type of reaction, text and tags of a comment, etc.) used to describe an action.*

Intuitively, a log tuple corresponds to an observation of $l.a$ performed by the user $l.u$ on a given object $l.o$ along with the associated attributes of the observation $\lambda_1, \cdots, \lambda_k$. If and action a_2 occurs after a_1 in a log, then the action a_2 occurred temporally after a_1.

Definition 4.1.2 (Activity Log)**.** *An* Activity Log *(L) is a finite sequence of log tuples related to a specific time interval* $]t, t+\Delta t]$

To properly analyze the actions made up by users on multimedia objects with respect to the time they are made, we provide the following definition of *Time Window Log*.

Definition 4.1.3. *(Time Window Log) Given a log* $L(t_i, t_f)$*, a Time Window Log* $w(t_1, t_2)$ *is a subset of L such that the timestamp of each action in* $w(t_1, t_2)$ *belongs to* $[t_1, t_2]$*, where* $t_1 \geq t_i$ *and* $t_2 \leq t_f$*.*

Once defined the concept of *Time Window Log*, it is possible to introduce the *Time Interval Analysis*, that is the time horizon for the required analysis.

Definition 4.1.4. *(Time Interval Analysis) Given a log* $L(t_i, t_f)$ *and two instances of time* t_s *and* t_e*, a* Time Interval Analysis $W(t_s, t_e) = w_0, w_1, ..., w_n$ *is the set of every* Time Window Log w_i*.*

Thus, analyzing the *Activity Log*, it is possible to build a sequence of MSNs, called *Multimedia Social Networks set*, in which each MSN is composed by nodes corresponding to users and multimedia objects and hyperedges representing each action made by users with respect to a specific time window.

Definition 4.1.5 (Multimedia Social Networks Set)**.** *Let* $W(t_s, t_e)$ *be a Time Interval Analysis, a Multimedia Social Networks set* $\mathcal{MSN}$ *is defined as*

$$\mathcal{MSN} = \{MSN(t_1), MSN(t_2), \ldots, MSN(t_n)\}$$

To analyze how the actions in each MSN are correlated, we have to define the concept of *relevant path*, corresponding to admissible sequences of hyperedges on which a user can spread her influence with respect

to another one. Note that, with abuse of notation, $v_i \underset{\tau}{\rightarrow} v_j$ iff $\exists u_i \in V_i \wedge u_k \in V_j : u_i \underset{\tau}{\rightarrow} u_j$.

Definition 4.1.6 (τ-Relevant path)**.** *A relevant path between two vertices v_{s_1} and v_{s_k} of a MSN is a sequence of distinct vertices and hyperedges $v_{s_1}, e_{s_1}, v_{s_2}, ..., e_{s_{k-1}}, u_{s_k}$ such that $\{v_{s_i}, v_{v_{i+1}}\} \subseteq V_{e_{s_i}}$ for $2 \leq i \leq k-1 \wedge$ $v_{s_i} \underset{\tau}{\rightarrow} v_{s_{i+1}}$.*

In the literature, several approaches have been proposed to model human behavior in On-line Social Network. Particular interest aroused the six principles of influence of Cialdini[29] which represent the fundamentals that describe how a user exerts influence on another one. In particular, exploiting the information contained in an MSN and the principles of reciprocity and liking, it is possible to provide the following definition of *reaction operator*, as shown in [9].

Definition 4.1.7 (Reaction Operator)**.** *The* Reaction Operator $reac^{\Delta t}(a_1, a_2)$ *between two actions a_1 of user u_i and a_2 of user u_j performed on the same object o (or on similar objects[1]) is the probability that a_2 occurs after a_1 within the interval Δt.*

The definition of *Reaction operator* is also based on the *consensus* principle representing the behavior of users that observe and analyze the actions of others to determine their own. In particular, the proposed model represents this facet through similarity relationships among multimedia objects, that can be computed using high and low level features.

Example 4.1.1 (Log and reaction operator)**.** *Consider a log, obtained from Flickr, whose associated sequence of actions is:*
⟨ *publishing, vinni, photo1,10/05/2017 13:30, 'sunset'* ⟩,

[1]The evaluation of such condition needs the definition of a similarity function between two objects.

⟨ *like, flora, photo1,10/05/2017 13:31* ⟩,
⟨ *comment, flora, photo1,10/05/2017 13:32, 'wonderful'* ⟩,
⟨ *publishing, picus, photo2,10/05/2017 13:38, 'sunset'* ⟩,
⟨ *like, giank, photo2,10/05/2017 13:40* ⟩,
⟨ *like, boscus, photo2,10/05/2017 13:42,* ⟩,
⟨ *comment, vinni, photo2,10/05/2017 13:45,''you too..'* ⟩
⟨ *like, boscus, photo1,10/05/2017 13:47* ⟩

It is simple to observe that considering $\Delta t = 2$ *minutes the reaction operator returns a probability value of* 1 *for the couple of actions (*publishing,like*). In turn, the probability value is* 0.5 *for the couple of actions (*publishing,comment*) and* 0.33 *for (*like,like*). If we consider a more wide temporal interval and assume the images published by* picus *and* vinni *very similar, the reaction operator will return a probability value of* 0.5 *for the couple of actions (*publishing,publishing*).*

Moreover, it is simple to observe that the following property stands for the reaction operator:

$$\begin{aligned} reac^{\Delta t}(a_1, a_2) > \tau^1 \wedge reac^{\Delta t}(a_2, a_3) > \tau^2 \\ \rightarrow reac^{\Delta t}(a_1, a_3) > \tau^3 \\ \tau^3 = f(\tau^1, \tau^2) \end{aligned}$$

f being a function whose value is less then $min(\tau^1, \tau^2)$ and $\tau \in [0, 1]$ being a probability value.

Human decisions are influenced by the growing overload of information that led us to identify shortcuts or rules of thumb. Understanding this phenomenon can be useful to improve significantly the way in which people are influenced.

By exploiting the influence operator and the consistency principle it is possible to define the following *influence operator*.

Definition 4.1.8 (Influence Operator)**.** *Let $u_1, u_2 \in U$ be respectively two users. We say that $u_1 \underset{\tau}{\rightarrow} u_2$, if each action $a_{u_1} \in A_{u_1}(t)$ of user u_1 at time t determines an action $a_{u_2} \in A_{u_2}(t, \Delta t)$ of user u_2 in the interval $]t, \ldots, \Delta t]$ within a log L:*

$$u_1 \underset{\tau}{\rightarrow} u_2 \iff \forall t_i \in T \; \exists a_1 \in A_{u_1}(t_i), a_2 \in A_{u_2}(t_i, \Delta t) \in L : reac^{\Delta t}(a_1, a_2) \geq \tau$$

$T = \{t_1, t_2, \ldots, t_m\}$ being a sequence of temporal instants such that $t_1 < t_2 < \ldots t_m$ and $\tau \in [0, 1]$ a probability value.

The *influence operator* estimates the influence of a user u_1 on u_2 within the time instance $t + \Delta t$.

Example 4.1.2 (Log and influence operator)**.** *Considering the log of Example 3.1 as past log (it can be used for the learning of reaction operator) and as current log the following sequence of actions:*
⟨ *publishing, vinni, photo1,11/05/2017 19:30, 'california'* ⟩,
⟨ *like, flora, photo1,10/05/2017 19:31* ⟩,
⟨ *like, boscus, photo1,10/05/2017 19:32* ⟩
⟨ *comment, flora, photo1,10/05/2017 19:32, 'wow!!!!!!!'* ⟩,
⟨ *publishing, picus, photo2,10/05/2017 19:58, 'hollywood'* ⟩,
⟨ *like, giank, photo2,10/05/2017 19:59* ⟩,
⟨ *like, boscus, photo2,10/05/2017 19:59* ⟩

It is simple to observe that considering $\Delta t = 2$ min we say that user vinni certainly influences flora and boscus and user picus certainly influences giank and boscus.

It is possible to compute the influence operator in several ways. The proposed approach allows to represent the probability of influence

between two nodes. In particular, exploiting the definition of *influence operator* and the action made by users u_i and u_j, it is possible to define the following two functions:

- *Reactivity of u_i with respect to u_j*: it corresponds to the ratio between number of reactions of u_i with respect to the action made by u_j ($r_{u_i u_j}$) and total number of reactions of u_i (a_{u_i}):

$$Reactivity_i = \frac{n^{r_{u_i uj}}}{n^{a_{u_i}}} \tag{4.1}$$

  This ratio represents how the user u_i is *reactive* with respect to u_j. In fact, it is easy to note that this value is equals to one iff the user u_i reacts on each content published by u_j;

- *Shareability of u_j*: it corresponds to the ratio between the number of reactions of u_i with respect to u_j ($r_{u_i u_j}$) against total number of actions of u_j (a_{u_j}):

$$Shareability_j = \frac{n^{r_{u_i uj}}}{n^{a_{u_j}}} \tag{4.2}$$

  This ratio represents how the user u_i reacted to the action of u_j, describing the capability of user u_j to influence directly user u_i.

Thus, the influence operator between two users can be computed in the following way:

$$\tau = (Reactivity_i * Shareability_j) = (\frac{n^{r_{u_i uj}}}{n^{a_{u_i}}} \times \frac{n^{r_{u_i uj}}}{n^{a_{u_j}}}) \tag{4.3}$$

It is possible to assume the following properties stand for the influence operator:

Property 4.1.1 (Not Self Reflexivity)**.** $u_1 \overset{\tau}{\nrightarrow} u_1$

Property 4.1.2 (Not Commutativity). $u_1 \underset{\tau}{\rightarrow} u_2$
$\nRightarrow u_2 \underset{\tau}{\rightarrow} u_1$

Property 4.1.3 (Not Transitivity). $u_1 \underset{\tau}{\rightarrow} u_2 \wedge u_2 \underset{\tau}{\rightarrow} u_3$
$\nRightarrow u_1 \underset{\tau}{\rightarrow} u_3$

Theorem 4.1.1 (Influence Diffusion). *Let u_1,u_2 and u_3 three users and L a given log,*

$$\begin{cases} u_1 \underset{\tau^1}{\rightarrow} u_2 \\ u_2 \underset{\tau^2}{\rightarrow} u_3 \end{cases} \Rightarrow u_1 \underset{\tau^3}{\rightarrow} u_3 \tag{4.4}$$

$\tau^3 \leq min(\tau^1, \tau^2)$.

Proof. Let us consider the definition of influence operator and the property of reaction operator; we observe that, for the theorem hypothesis, there will always exist the two actions $a_{u_1} \in A_{u_1}(t)$ and $a_{u_3} \in A_{u_3}(t : \Delta t')$ and the reaction operator will always return for the couple (a_1, a_3) a probability $\tau^3 \leq min(\tau^1, \tau^2)$. □

Analyzing the interaction between users, it is possible to note that multimedia content plays an increasingly important role. Using the previous definition of *influence operator* it is possible to define two types of influence in the MSN. The first one is related to two users that interact on the same multimedia contents while in the second type the influence between two users is based on the similarity between pairs of multimedia contents. Formally, it is possible to provide the following definitions.

Definition 4.1.9 (Direct Influence). *Let $MSN = (V, H)$, $S \subseteq H$ and u_i,u_j be respectively an influence graph, a set of similarity relationship and two users, a user u_i* directly influences *u_j if there exists a τ-relevant path $rp = (h_1, \ldots, h_n)$, with $h_1, e_2, \ldots, h_n \in H \setminus \{S\}$, that connects u_i to u_j.*

Definition 4.1.10 (Indirect Influence)**.** *Let $MSN = (V, H)$ and u_i,u_j be respectively an influence graph and two users, a user u_i* indirectly influences *u_j if there exists a τ-relevant path $rp = (h_1, \ldots, h_n)$, with $h_1, e_2, \ldots, h_n \in H$ that connects u_i to u_j.*

Exploiting the previous definitions of influence, it is possible to compute the *Influence graph*, an homogeneous graph, in two stages. Firstly, the direct influence between users is computed on each MSN to provide a first approximation about influence weights between users. Successively, the second stage considers only paths, included similarity hyperedges, that interconnect pairs of users with an influence weight, computed in the first stage, greater than a given threshold Θ. This concept led to represent the human behavior to emulate with greater likelihood the actions made by social ties with respect to other people.

Thus, the *Influence graph* is an homogeneous graph that describes how a user exerts its influence on others. Formally, it is possible to define the *Influence graph* in the following way.

Definition 4.1.11 (Influence Graph)**.** *An* Influence Graph *is a labeled graph $G = (V, E, \tau)$ where:*

- *V is the set of nodes such that each $v \in V$ corresponds to a user $u \in U$;*
- *$E \subseteq V \times V$ is the set of edges (with no self-loops);*
- *$\tau : V \times V \to [0,1]$ is a function that assigns to each edge $e = (v_i, v_j)$ a label, representing the probability that user u_i can influence user u_j.*

Thus, the resulting outcome is a sequence of *Influence Graphs*, as shown in figure 4.1 that are retrieved by computing the *Influence Operator* on each MSN. This sequence of *Influence Graph* is stored into a Data Warehouse, called *Graph Warehouse*.

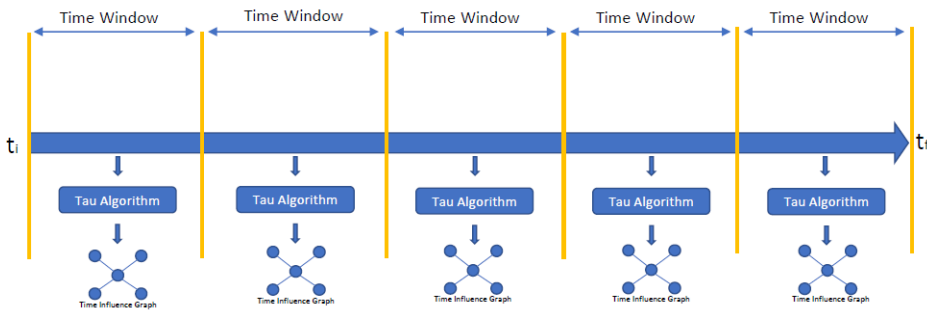


Figure 4.1: Sequence of Influence graph.

Definition 4.1.12 (Graph Warehouse)**.** *A Graph Warehouse is a collection of vertices' relationships concerning to a sequences of influencing relationships between users included in the examined* Influence Graph *to analyze information for supporting different applications.*

In particular, note that a Graph Warehouse has all the main characteristics of a traditional Data Warehouse:

- *Subject Oriented*: it is oriented to the users of the examined network;
- *Integrated*: it is built using data retrieved by different types of MSNs;
- *Time Varying*: it is based on time, in fact it considers different graph for each time window.
- *Not Volatile*: once time *influence graphs* are computed, they are not changed.

Using the information contained in a Graph Warehouse, it is possible to compute the influence among two users over different temporal windows, analyzing the changes of the network, in several way. This approach is based on the seventh principle of Cialdini [30], called *unity*, that allows

to consider a shared identity between the influenceers and influenced. Figure 4.2 shows a synthetic schema of the proposed approach.

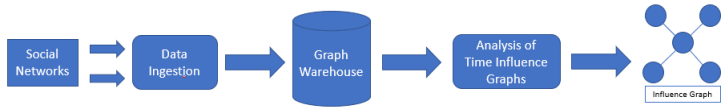


Figure 4.2: Example of Graph Warehouse

In the following several algorithms have been proposed to deal with the problem of Influence Maximization (IM).

4.1.2 Influence Maximization Algorithm

In this section several algorithms, based on our *Influence Model* in order to deal with the Influence Maximization (IM) problem, are proposed.

IM Algorithm based on Stochastic approach

In this section it is described an algorithm [9] based on the stochastic approach for the *Influence Maximization* problem. This approach leverages multimedia objects that represent a means of interaction among users; for instance it is possible to estimate the influence that a user exerts on another one or define communities of users analyzing *relevant paths*.

This approach is an extension of sketch based methods [136, 134], in which several samples are computed for identifying the nodes' set covering a large number of samples to infer an approximation of solution. Each sample is the set of nodes reachable from a given vertex on the graph, obtained by removing an edge with a given probability.

The idea behind this algorithm is to reduce the required number of samples exploiting the multimedia objects relevance. In particular, the following three matrices have been defined.

Definition 4.1.13 (User-Relevant Path Matrix)**.** *Let U and P be respectively the sets of users and relevant paths of MSN, a* User-Relevant Path Matrix *is defined as:*

$$\boldsymbol{UP} = \{up_{u_i p_j}\} = \begin{cases} \tau & u_i \, influences \, on \, path \, p_j \\ 0 & else \end{cases}$$

where τ corresponds to the probability that u_i exerts its influence over path p_j

Definition 4.1.14 (Relevant Path-Multimedia Object Matrix)**.** *Let M and P be respectively the sets of multimedia objects and relevant paths of MSN, a* Relevant Path-Multimedia Object Matrix *is defined as:*

$$\boldsymbol{PM} = \{pm_{p_i m_j}\} = \begin{cases} 1 & m_j \, belongs \, to \, the \, path \, p_i \\ 0 & else \end{cases}$$

The above defined matrices can be used to compute a *User-Multimedia Objects Matrix* considering the number of paths among vertices considering the effects of influence operator.

Definition 4.1.15 (User-Multimedia Object Matrix)**.** *Let U, M, $\boldsymbol{UP}$,$\boldsymbol{PM}$ be respectively the sets of users, multimedia objects and user-relevant path and relevant path-multimedia objects matrices, the* User-Multimedia Object Matrix *is defined as:*

$$\boldsymbol{UM} = \{um_{u_i m_j}\} = \sum_{u_i \in U} \Big(\prod_{p_k \in P} up_{u_i p_k} * pm_{p_k m_j} \Big)$$

In other words, the idea is to reduce the amount of required samples choosing the minimum number of relevant multimedia objects that allows us to identify an appropriate number of samples. Eventually, the seed set is computed as vertices collection that maximize the fraction of samples covered from each node.

Algorithm 2 τ-Diffusion algorithm

procedure τ-DIFFUSION ALGORITHM(MSN, τ^*, θ,k)
    - - Input: MSN (the hypergraph of MSN)
    - - Input: τ^* (the threshold of influence)
    - - Input: θ (the threshold of MOs relevance)
    - - Input: k (the desired size of seed set)
    - - Output: S (the seed set)
    - - Temp: **UP** (the U-RP matrix)
    - - Temp: **PM** (the RP-MO matrix)
    - - Temp: **UM** (the U-MO matrix)
    - - Temp : sum (the relevance of given MO)
    - - Temp : E (the set of examined relevant path)
    - - Temp : SS (the set of samples)
    - - Temp : T (the vertices set for sampling generation)
    UP $\leftarrow$ compute_matrix_UP(MSN,τ_s)
    PM $\leftarrow$ compute_matrix_PM(MSN,τ_s)
    UM $\leftarrow$ compute_matrix_UM(MSN,τ_s)
    E $\leftarrow$ 0
    T $\leftarrow$ 0
    SS $\leftarrow$ 0
    j $\leftarrow$ argmax$_{u_i}\{\sum_{m_j} um_{u_i m_j}\}$
    sum $\leftarrow$ compute_sum(UM,j)
    if $(sum \geq \theta)$ **then**
        E $\leftarrow \{p_i : pm_{p_i m_j} = 1\}$
        T $\leftarrow \{\forall e \in E, u_i : up_{u_i p_e} \neq 0\}$
    end if
    SS $\leftarrow$ Sample_generation (T,**UP**)
    S $\leftarrow$ Node_covering(SS,k)
end procedure

The τ^* value is chosen to prune the sequences of actions on which the influence among users is less than τ^*. In the Algorithm 2, the attention is focused on the relevance of multimedia objects involved in the more relevant paths through *compute_sum* function. In this way, it is identified a set of vertices from which generate the samples that are exploited by the *Sample_generation* and *Node_covering* function to find the desired subset.

Example 4.1.3. *To better explain the proposed idea, let $Vinni$, $Giank$, $Picus$, $Flora$, $MisterX$ and $MisterY$ and MO_1, MO_2 and MO_3 be respectively six users and three images, as described in the following log.*

⟨ *publishing, Vinni,* MO_1*,10/05/2017 19:30, 'california'* ⟩,
⟨ *comment, Flora,* MO_1*,10/05/2017 19:31, 'wow!!!!!!!'* ⟩,
⟨ *like, MisterX,* MO_1*,10/05/2017 19:32* ⟩
⟨ *like, MisterY,* MO_1*,10/05/2017 19:33* ⟩
⟨ *publishing, Vinni,* MO_2*,10/05/2017 19:35, 'beach'* ⟩,
⟨ *like, MisterX,* MO_2*,10/05/2017 19:36* ⟩
⟨ *like, MisterY,* MO_2*,10/05/2017 19:37* ⟩
⟨ *publishing, picus,* MO_3*,10/05/2017 19:58, 'sunset'* ⟩,
⟨ *like, Giank,* MO_3*,10/05/2017 19:59* ⟩,
⟨ *like, MisterX,* MO_3*,10/05/2017 19:59* ⟩

Analyzing the activity log, we build the hypergraph representation of MSN, showed in Figure 4.3.

In the first stage, the proposed algorithm compute the ***UP***,***PM*** *and* ***UM*** *matrices based on the given value of threshold τ^*; in particular in the following we show the obtained matrices by choosing τ^* equal to* 0.25.

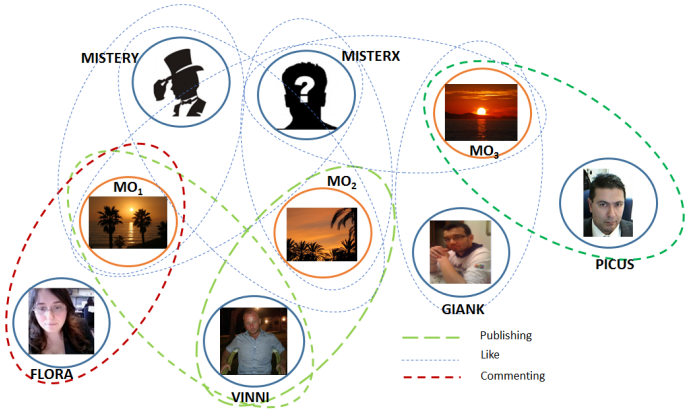


Figure 4.3: Example of MSN

$$\begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.0 & 1.0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.33 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} =$$

$$\begin{bmatrix} 3.0 & 2.0 & 0 \\ 0 & 0 & 2.0 \\ 0 & 0 & 0.33 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Thus, analyzing the generated matrices, we can compute the number of required samples based on multimedia objects relevance, corresponding to the sum of each column of ***UM*** *matrix, and the value of user selected*

threshold θ. In this example, it is easy to note that we take into account only the vertices Vinni *and* Picus *for the generation of the samples, choosing a threshold value equals to* 2.3*. These two vertices corresponds to the users that allow us to maximize the spread on the network, exerting influences on a large number of other users.*

IM algorithm based on game theory

Viral marketing exploits the features of MSNs for leading customers to share product information with their friends. In Influence Analysis, one of the main problems is to maximize the number of people that become aware of a given product, finding the best set of users to start the diffusion and maximize the spread. Thus, the problem is to find the seed set only considering the structure of the social network without any information about influence probabilities. For this reason, it is proposed a *Combinatorial Multi-Armed Bandit (CMAB)*[19, 131, 144] approach on MSNs that allows us to estimates the influence probabilities.

The Combinatorial Multi-armed bandit (CMAB) is an algorithm of influence maximization based on rounds. This approach uses a Multimedia Social Network without unknown distribution of influence probabilities with the aim to minimize the regret, corresponding to the difference between the spread of influence and the optimal solution. Thus at each step, the algorithm attempts to reduce the regret and at the same time to improve our knowledge of the influence probabilities distribution. The regret ρ after T round is described by the following formula 4.5, that represents the difference between suboptimal strategy S^* (knowing the probability) and the CMAB strategy S_s.

$$\rho = T\sigma(S^*) - \sum_{s=1}^{T} \bar{\sigma}(S_s) \tag{4.5}$$

Two types of action can be made using CMAB algorithm :

CMAB	Symbol	IM
Base arm	i	edge $(u.v)$
Reward arm i in round s	$X_{i,s}$	Status Relevant path (u,v) (*active/inactive*)
Mean of distribution arm i	μ_i	Influence probability $p_{u,v}$
Superarm	A	Outgoing edge E_S from nodes $u \in S$
Reward in round s	r_s	Spread σ in the s^{th} IM attempt
No. of times i is triggered in s rounds	$T_{i,s}$	No. of times u becomes active in s diffusions

Table 4.1: Mapping of the CMAB framework to IM

1. Exploration used to improve the knowledge learning the influence probabilities in the network;

2. Exploitation used to led to a large spread exploiting the knowledge obtained;

In table 4.1, it is shown the mapping that highlights the similarities between IM and CMAB.

The proposed framework considers m arms each of which has an unknown reward distribution. The process evolves into a fixed number T of rounds, in each round s an arm is played and is observed the reward generated. Moreover, in each round s it is adopted exploration-exploitation trade-off, that allows to minimize the regret (playing the arms that are considered considered to be better) and improve the knowledge (playing arms never used or less used).

The Combinatorial Multi-Armed Bandit paradigm extends the previous paradigm introducing a concept of *superarm* (a set of arms). In each round is introduced the use of an approximation oracle to find the best (super)arm to play (Gai et al.[52] and Chen et al. [28]). The superarm A can trigger others arms - (Chen et al. [28]), in this case the reward obtained could be a non-linear combination of the single rewards.

The proposed framework considers m arms each of which has a random variable $X_{i,s} \in [0,1]$ with mean μ_i, which represents the reward obtained when the arm i is triggered in the round s. The CMAB paradigm plays a superarm A in each of the T round, in this way all arms in A are triggered (this happens with probabilities $p = 1$). Playing the superarm A it is possible trigger others arms; it is possible to define p_A^i the probability of arm i to be triggered when superarm A is played. In each round it is used an approximation oracle to find the best (super)arm to play; the oracle works on the current means $\vec{\hat{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_m)$. After that, the superarm is played and consequently several arms i are triggered. Finally, it is possible to note that the reward obtained from each arm allows to update their mean estimate $\hat{\mu}_i$. Thus, the number of times that an arm i has been triggered until the round s is defined as $T_{i,s}$. The process continues until $s = T$.

In the Algorithm 3 it is described the CMAB approach, that receives in input an influence graph $G = (V, E)$, without knowledge of probabilities of influence between two users, in which the set E is composed by the edges included in the identified relevant path while the set of vertices corresponds to the collection of entities involved in relationship of E. At the beginning (row 2) it is used the vector $\vec{\hat{\mu}}_0$ that assign 0 or a low probability of influence to each relevant path (u, v). In each round the attempt is to reduce the regret through an *exploration-exploitation trade-off*. This latter is chosen through a flip coin.

When the chosen strategy is *EXPLOIT*, an influence maximization algorithm A (in our case TIM^+) is used to represent the approximation oracle. A seed set S with $|S| = k$, that is the Super-Arm SA, is chosen by solving an IM problem on MSN H with influence probability estimates $\vec{\hat{\mu}}$. Instead, when the chosen strategy is *EXPLORE*, it is selected a seed set S randomly.

Once obtained the Superarm SA, it is played starting the diffusion process that lead to active several inactive nodes and triggering other arm. At the end of the process it is valued the reward considering the number of nodes that are activated through Spread-Pregel feedback. This information is used to update the mean estimate $\vec{\hat{\mu}}$ by the formula 4.6, where $X_{i,s}$ represents the number of nodes activated playing the arm i until step s and $t \in [1, T]$.

This process allows to improve the knowledge in order to minimize the regret in the next step.

$$\hat{\mu}_i = \frac{\sum_{s=1}^{t} X_{i,s}}{T_{i,t}} \tag{4.6}$$

Algorithm 3 CMBA's algorithms for Influence Maximization without influence probabilities ($MSN(V,E)$, $\vec{\hat{\mu}}_0$, k, P_{tr}, T, ϵ, *Algorithm* TIM^+, *Cascade Mechanism* C)

```
1: procedure CMAB
2:     Initialize μ⃗̂ ← μ⃗̂_0
3:     ∀_i initialize T_i = 0
4:     for s = 1 → T do
5:         Flip coin and set is-exploit
6:         if is-exploit then
7:             SA = EXPLOIT(G, μ⃗̂, TIM^+(ϵ), k)
8:         else
9:             SA = EXPLORE(G, k)
10:        end if
11:        Play the Super-Arm SA and trigger other Arm with P_tr
12:        μ⃗̂ = UPDATE(C)
13:    end for
14:    return SA
15: end procedure
```

As it is possible to note in Figure 4.4, the process evolves in discrete step. In the first step (Figure 4.4(a)) we do not have information of

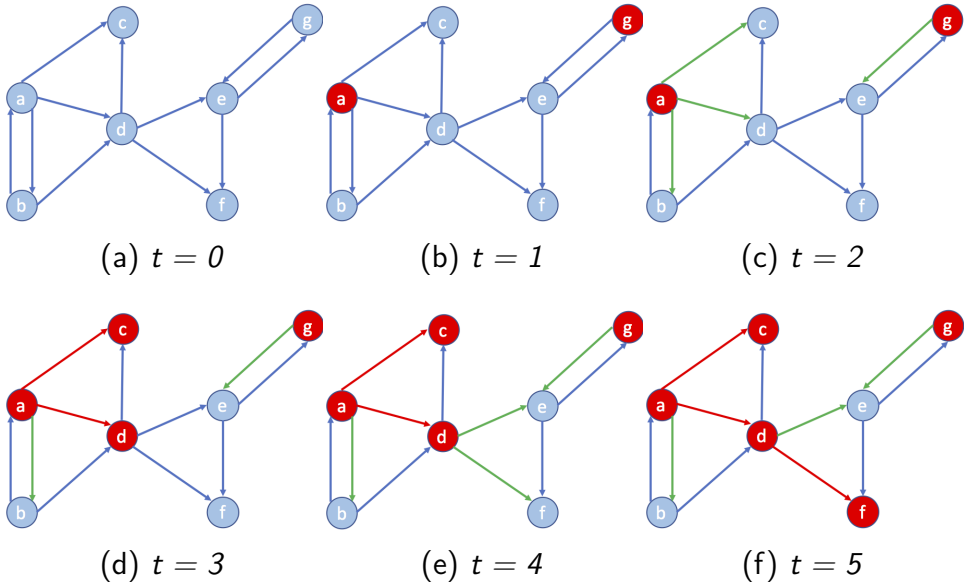


Figure 4.4: An example of triggering of a Super-Arm

influence. In (Figure 4.4(b)) the Super-Arm $SA = (a, g)$ is played, it triggers the relevant paths that attempt to activate the inactive nodes. In (Figure 4.4(d)) the nodes c and d become active . At the end of the process is observed the efficiency of the arms, this information is used to update the mean estimate.

To observe the success or failure of activation attempts it is adopted a new method that evaluates the status of each node. Once the node is influenced it is possible to keep track of the edge (the arm) that attempts to activate. In this way it is possible to update the mean estimate for each relevant path knowing the success/failure of this last.

A bio-inspired influence maximization algorithm

A biologically inspired approach for IM is the *ABC* algorithm [157] based on the bees' behaviors [122] within an hive. In particular, the ABC

algorithm mimics the collective foraging behavior of honey bees when searching food. Bees that found a highly profitable food source go to an area in the hive called the *dance floor*, and perform the *waggle dance*. Through the waggle dance, a scout bee communicates the location of its discovery to other bees, that can join the exploration of the flower patch. Since the length of the dance is proportional to the scouts rating of the food source, more bees get recruited to harvest the best rated flower patches. Thus, the algorithm is based on the combination of global-local search feature and the waggle dance process. The waggle dance represent the way that each bee can exploit to attract (influence) other bees and can correspond to a variety of user actions (e.g., publishing a content, posting a review, etc.) in MSNs. To represent the spread model over the net it is used a *Linear Threshold* (LT) model [77], where the "activation" of each node is based on the behavior of its neighbors.

More in details, the ABC algorithm works according to two fundamental steps: i) an initial user ranking is performed (to determine the most suitable employer bees to lead the food search campaign together with a set of scouts represented by their neighbors in the network); ii) a top-k selection of the most influential users within the initial set (represented by the employer or scout bees that are effective leaders on the base of their waggle dance) is carried out in an iterative manner.

However, the proposed method has the following limitations: it is not suitable for a generic OSN (being designed for Twitter) and the output node set could be a local maximum. Thus, we introduced some changes to the origin algorithm: we compute the k most influent nodes using the influential paths and the related influence graph as defined in the previous section. We assume that a waggle dance was successfully performed by a user u_i on a user u_j if an influential path has been instantiated between u_i and u_j. Furthermore, to solve the local maximum issue, we exploited a *tabu search* technique thus in each iteration, if the number of

Algorithm 4 Modified ABC algorithm

```
1: procedure IM-ABC(IG,k,N_s,ω_t)
2:     - - Input: IG (the influence graph)
3:     - - Input: k (the number of influentials)
4:     - - Input: N_s (the number of neighbor scouts)
5:     - - Input: ω_t (the diffusion threshold)
6:     - - Output: Emp (the set of the employer bees)
7:     - - Output: fit (the employers' fitness values)
8:     - -Temp : Scouts (the set of scouts)
9:     - -Temp : fit_s (the fitness of a scout)
10:    Emp ← top-k ranked IG nodes
11:    fit ← fitness(IG ,Emp,ω_t)
12:    while (termination condition is not met) do
13:        Scouts ← neighborhood(IG,Emp,N_s)
14:        if (Scouts=∅ ∨ | Scouts| < N_s) then
15:            Add to Scouts a set of random nodes
16:        end if
17:        for (each s ∈ Scouts) do
18:            fit_s ← local_fitness(IG,Scouts,ω_t)
19:            if (∃ e ∈ Emp: fit_s > fit(e) ) then
20:                Emp=Emp-e
21:                Emp=Emp ∪ s
22:                mark e as onlooker bee
23:            end if
24:        end for
25:        Update Scouts
26:    end while
27:    return Emp
28: end procedure
```

scouts is not equal to the user defined parameter, the algorithm chooses a random number of scouts. In addition, we can leverage several centrality measures for the ranking stage.

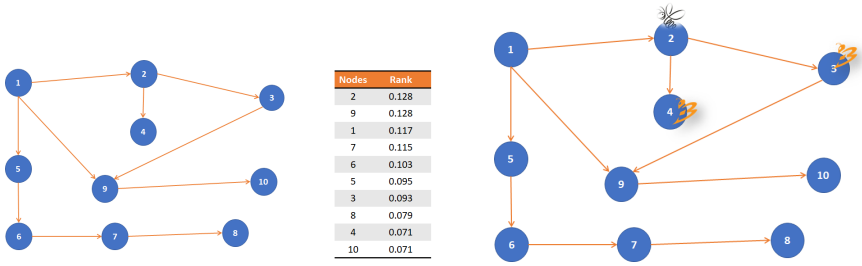
Algorithm 4 summarizes the described influence maximization process.The *employer bees*, who are used to locate influential opinion leaders in the network, are initially assigned to the top-k nodes from the node ranking process. The *scout bees* are used to explore the nearest neighbor nodes of employer bees for better solutions, while the *onlooker bees* indicate the users influenced by some influential users. In each iteration the local_fitness is calculated by the maximum number of the influence graph nodes that can be influenced (reachable in a single step) by a single node with a probability value grater that ω_t (in according to the LT model).The global_fitness values is maximum number of nodes that can be influenced by a group of k nodes.

Example 4.1.4. *Let IG be the influence graph shown in figure 4.5a and k and N_s be respectively equal to 1 and 2. A first set of employers is composed only by the node 2 using the PageRank algorithm, as shown in figure 4.5a. Successively, the nodes 3 and 4 are chosen as scouts using nearest neighbors of initial employer bees. Thus, an iterative process starts updating the status of bee to employer and the employers to an outlookers if a scout bee has a fitness*[2] *value greater then each employers bee. In this example, the proposed algorithms returns the node 2 as seed set because it has fitness value greater than the scout bees producing the outcome as it is possible to note in figure 4.5b.*

4.2 Community Detection

Nowadays we live in an increasingly global and connected context, in which each person can share ideas, feelings, etc. with other people in the

[2]number of nodes that can be reached in a given number of steps



(a) *First Step* (b) *Second Step*

Figure 4.5: An example of bio-inspired IM algorithm.

world. In this way, a large amount of heterogeneous data are generated. Mining communities allows to facilitate the analysis of networks.

For this reason, in this section, the proposed approach for community detection, that exploits the features of Multimedia Social Network model, is described.

4.2.1 Proposed Algorithm

In this section we describe an approach for community detection on Multimedia Social Networks. The idea behind our approach is to identify groups of people based on the analysis of actions made by users on multimedia objects.

Example 4.2.1 (Example of network)**.** *In figure 4.6 we show an example of MSN with seven users and multimedia objects. How it is possible to see each hyperedge corresponds to an action made by a user on the network; for instance u_5 user made two types of actions first one is published of two multimedia contents M_5 and $M6$ and comment M_3.*

In our model there are several paths that allow to connect two nodes on which two users can perform several action on Multimedia Social Network; for instances they can publish multimedia content or tagging a

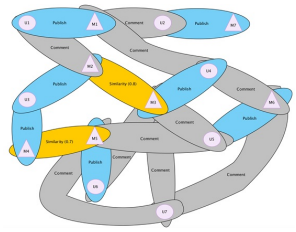


Figure 4.6: An example of MSN

friend for a specific event or make a comment on another multimedia content provided by another user and so on.

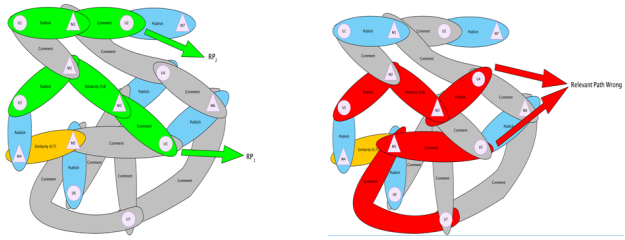
Definition 4.2.1 (Social Path)**.** *A* social path *between two vertices* v_{s_1} *and* v_{s_k} *of an OSN is a sequence of distinct vertices and hyperedges* $sp(v_{s_1}, v_{s_k}) = v_{s_1}, he_{s_1}, v_{s_2}, ..., he_{s_{k-1}}, v_{s_k}$ *such that* $\{v_{s_i}, v_{s_{i+1}}\} \subseteq HV_{he_{s_i}}$ *for* $1 \leq i \leq k-1$*. The* length γ *of the hyperpath is* $\alpha \cdot \sum_{i=1}^{k-1} \frac{1}{\omega(he_{s_i})}$*,* α *being a normalizing factor. We say that a social path* contains *a vertex* v_h *if* $\exists he_{s_i} : v_h \in he_{s_i}$*.*

Social paths between two nodes leverage the different kinds of relationships: a given path can “directly” connect two users because they are “friends" or members of the same group, or “indirectly”, as they have commented the same picture.

To analyze how the actions made on Multimedia Social Network from users are correlated, we have to define the concept of relevant path, corresponding to admissible sequences of hyper-edge on which a given user can spread his idea respect to another one.

Definition 4.2.2 (Relevance Social Path)**.** *Let* Θ *be a set of conditions defined over the attributes of vertices and hyperedges, a* relevance social path *is a social path satisfying* Θ*.*

A particular kind of relevance social path is constituted by the *influential paths* that connect two users and by which a user can “influence”



(a) Relevant path (b) Relevant path

Figure 4.7: An example of Influence path

other users. As an example, in Flickr a user u_i influences another user u_j, if u_j adds to her/his favorites any photo of u_i, or if u_j positively comments a photo (or one similar to) that u_i has just published. In Twitter, the influence is mainly related to the re-tweet actions, thus the user u_i influences the user u_j, if u_j has re-tweeted any tweet of u_i. Similarly in Yelp, the user u_i influences u_j, if the user u_j posts a review of the same sentiment of the review previously posted by u_i on the same business object. Indeed, the type of influential paths that can be considered depends on the Social Network and on the analytical goals.

Concerning the Flickr example, all the influential paths have the form $u_1, he_1, o_1, he_2, u_2$, where $he_1.\mathsf{type} = \text{“}publishing\text{”} \wedge he_2.\mathsf{type} = \text{“}add\_favorite\text{”} \wedge (he_2.\mathsf{time} - he_1.\mathsf{time}) \leq \Delta t$ constitute the set of conditions Θ, being Δt a given time.

Example 4.2.2 (Influential Path)**.** *Considering the MSN shown in example 4.2.1, we can identify several social paths. The relevance social path is computed by choosing some paths based on specific rules. In particular, in the figure 4.7a taking in account the following relevant path publish similarity content we have the following path while in the second one (fig 4.7b) publish and comment.*

Thus, we analyze the interactions among users exploiting the information of the relevant paths that are represented in a square matrix, called Weighted Relevant Path matrix.

Definition 4.2.3 (Weighted Relevant Path Matrix)**.** *Let MSN be a multimedia social network, RP a list of relevant path we define Weighted Relevant Path Matrix WRP as a square matrix $|U| \times |U|$, where U is the set of users in MSN. Each element of W is equal to*

$$WRP = \{wrp_{ij}\} = \begin{cases} \sum_{p \in RP} w_p \; if \;\; i \neq j \\ 0 \quad if \;\; i = j \end{cases}$$

where w_p corresponds to the weight of each relevant path established between two users i e j.

Example 4.2.3 (Example of Weighted Relevant Path)**.** *In this example we compute the weighted matrix considering the influence paths shown in figure 4.7a. Leveraging the information of relevant path, we can compute the element 3-5 of Weighted Relevant matrix by the sum of product of two relevant paths that interconnect these two users, as shown in figure 4.8.*

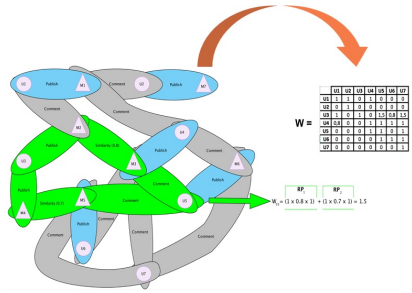


Figure 4.8: An example of WRP computation

Multimedia objects represent a new means of interaction among users; for instance it is possible to estimate the influence that a user exerts

on another one or define communities of users analyzing relevant paths. In particular the similarity relationships among multimedia contents provide new paths to spread idea or concepts in a MSN. Furthermore, we compute a symmetric matrix from the Weighted Relevant Path Matrix to properly manage the expansion phase.

Definition 4.2.4 (Weighted Users Matrix)**.** *Let WRP be a Weighted Relevant Path Matrix, we define a Weighted User Matrix WU as a square matrix $|U| \times |U|$ obtained by $WRP \times WRP^{\mathrm{T}}$*

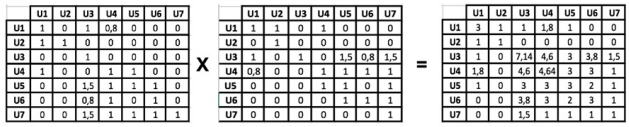


Figure 4.9: An example of Weighted User Matrix for a MSN

Example 4.2.4 (Example of Weighted User Matrix)**.**

Our approach, based on the methodology proposed in [64], is based on a vertex selection strategy that guarantees high coverage and good conductance on expansion of community. The main contributions are related to: (i) the model, that integrates both information about users of one or more OSNs and the content related to it generated and shared by a hypergraph data structure; (ii) the use of relationships based on similarity between two multimedia objects that allow to establish other paths; (iii) a new way to build the matrices used by the examined algorithm. In the first step, the algorithm identifies the nodes with highest weight degree as the seed nodes. The weight degree of node u is computed as the sum of column related to user u of Weighted User Matrix. In particular the algorithm is composed by two steps. Firstly, we choose as starting node the vertex v corresponds to $\arg\max_{u \in U}\{\sum_{v \in U} WU_{vu}\}$. Successively, we use the conductance measure to evaluate the quality of community during

the expansion phase; in fact the increase of nodes leads to include in the examined community people who are only marginally interested in shared contents, that corresponds to a decrease of conductance value.

Formally, the conductance is defined as $\phi(C_i) = \frac{cut(C_i)}{min\{deg(C_i), deg(\bar{C}_i)\}}$, where where $cut(C_i)$ denotes the size of cut induced by C_i, $\bar{C}_i$ the complement set of C_i and sum of degrees vertices in C_i

Once a seed node is identified, we perform an incremental expansion of examined community allowing to include a user that maximize the decrease value of conductance. This is an iterative process to define a community that ends when the conductance difference does not assume negative value.

Algorithm 5 Community detection algorithm

```
1: procedure MSN_Community_Detection(MSN)
2:     C ← ∅
3:     Compute_Matrix WRP
4:     Compute_Matrix WU
5:     while more_visited_nodes do
6:         C_i = ∅
7:         u ← arg max_{u∈U} {Σ_{v∈U} WU_vu}
8:         C_i ← C_i ∪ {u}
9:         while (φ(C_i) − φ(C_i^`) ≥ 0) do
10:            u ← arg max_{u∈U} {Σ_{v∈U} WU_vu}
11:            C_i ← C_i ∪ {u}
12:        end while
13:        C ← C ∪ C_i
14:        i ← i + 1
15:    end while
16:    return C
17: end procedure
```

Example 4.2.5 (Example of Approach)**.** *In this example we show how our approach allows to detect community in the network of the example*

	Incremental expansion	Conductance score
Iteration 1		
Seed 1	3	0,661
	3, 4	0,581
	3, 4, 6	0,495
	3, 4, 6, 5	0,427
Community 1	3, 4, 5, 6	
Iteration 2		
Seed 2	1	0,615
	1, 2	0,49
Community 2	1, 2	
Iteration 3		
Seed 3	7	0,818
Community 3	7	

(a) Relevant path

	Incremental expansion	Conductance score
Iteration 1		
Seed 1	3	0,661
	3, 4	0,581
	3, 4, 6	0,495
	3, 4, 6, 5	0,427
Community 1	3, 4, 5, 6	
Iteration 2		
Seed 2	1	0,615
	1, 2	0,49
Community 2	1, 2	
Iteration 3		
Seed 3	7	0,818
	7, 3	0,637
	7, 3, 4	0,554
	7, 3, 4, 6	0,465
	7, 3, 4, 6, 5	0,394
Community 3	3, 4, 5, 6, 7	

(b) Relevant path

Figure 4.10: An example of Community detection

4.2.1. In figure 4.10 we analyze the obtained communities using the proposed community detection approach. We obtain that user with id 7 is included in the large community comparing the figure 4.10a and 4.10b.

5 Evaluation

In this section the proposed diffusion models based on the specific social network with a real dataset have been evaluated. The goal of the section is to show that the models have accurate propagation probabilities reflecting the *social behavior* of people in a Social Network. The experiments have done comparing IM algorithms using the proposed models and other state-of-the-art ones.

We also describe the used experimental architecture, based on Big Data technologies and designed for developing reliable, scalable and completely automated data pipelines. In particular, this infrastructure includes handle mechanisms for ingesting, managing and processing large amount of data. This infrastructure is based on *Lambda* architecture allowing to handle data by obtaining advantage of both batch and stream processing methods.

Eventually, we show that our model adapts itself on the data, especially if the social *actions-reactions* increases. We have considered two different experimental protocols, the first without similarities among images and the second considering similarities. We have demonstrated that our model is more efficient than the other approaches described in the literature.

5.1 Platform Architecture

The proposed architecture is composed by two macro components: the first one, on the left of Figure 5.1, has the aim to crawl, clean and store information of several On-line Social Networks while the second one is based on a Spark cluster infrastructure for supporting several applications. In the next section more details about these two components are provided.

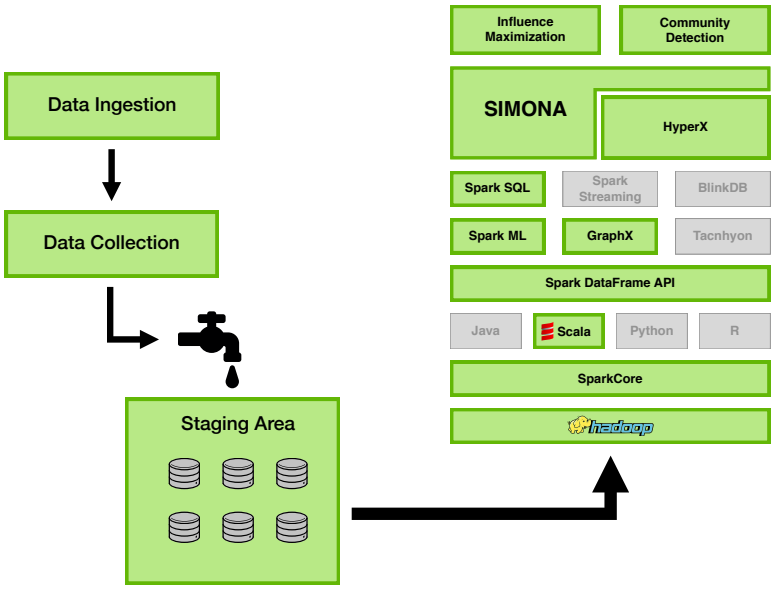


Figure 5.1: Big Data Infrastructure

5.1.1 Ingestion & Data Collector modules

The ingestion module provides the capabilities to obtain a direct interface with data sources. It is the first layer of the Big Data pipeline for

managing data coming from several sources at variable speed that can be directly used or stored. In particular, it is composed by different connectors that allow to crawl data from several social networks.

Successively, Data collector module manages the transportation data from ingestion layer to the staging area. It acts as a mediator based on a messaging system between these two modules. It lets to properly manage and store streams of records in distributed and a fault-tolerant way.

5.1.2 Staging area module

The Staging area module is used to store streams of records (actions) made by different users on several On-line Social Networks. This module allows to perform several essential tasks (data aggregation, cleaning, or merging etc.) prior to process them in the processing phase.

In this architecture, a column oriented NoSQL database is chosen because it is properly used for handling huge amount of data and for performing aggregation operations. In particular, this type of database adopts a column-based approach for storing the crawled streams of record which are properly suited for storing information contained in logs crawled from OSNs, as described in the 4.1.1.

5.1.3 Processing modules

The second stage, on the right of Figure 5.1, has the aim to support several applications (i.e. community detection, influence analysis, lurker detection etc.).

Multimedia Social Network Analytics engine combines the following three layers: a distributed file system based on HDFS, a data processing layer, that use HyperX library and Spark modules, and *SIMONA*(SocIal and Multimedia On-line Network Analysis).

Initially, data stored in the staging area are filtered according to the application requirements and they are loaded on *Hadoop HDFS*, a distributed file system designed to run on commodity hardware.

On top of HDFS there is the data processing engine based on *Apache Spark* that allows to handle properly the data model described in chapter 3 using the *API Graphx* and *MLlib*, and *HyperX* for performing distributed computation.

Apache Spark, whose architecture is shown in Figure 5.2, is a cluster computing platform used to analyze data across a cluster of commodity computers. Spark is a unified engine that provides APIs for programming entire clusters including support for SQL queries, streaming data, machine learning and graph processing. It supports several types of workloads like batch applications, iterative algorithms, interactive queries and streaming.

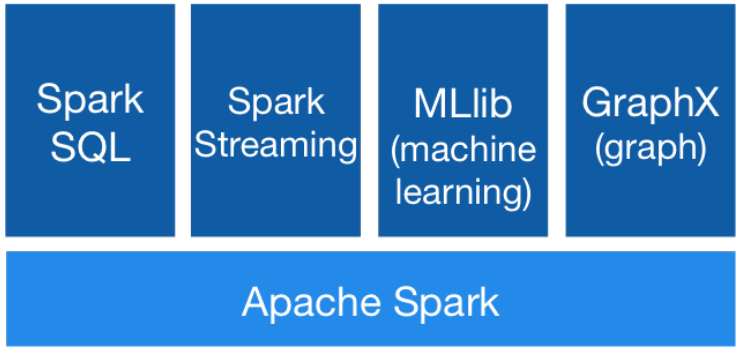


Figure 5.2: Architecture of Apache Spark

On the top of Spark we use HyperX [67] library that extends the GraphX module's capabilities, for manipulating and processing the MSN data model. HyperX, whose architecture is shown in Figure 5.3, is a thin layer built on Spark HyperX to address the dimensioning problem by operating directly on a native hypergraph rather than on a converted graph. Specifically, it stores hyperedges and vertices using two

Resilient Distributed Datasets [1] (RDD): *vRDD* for vertices and *hRDD* for hyperedges.

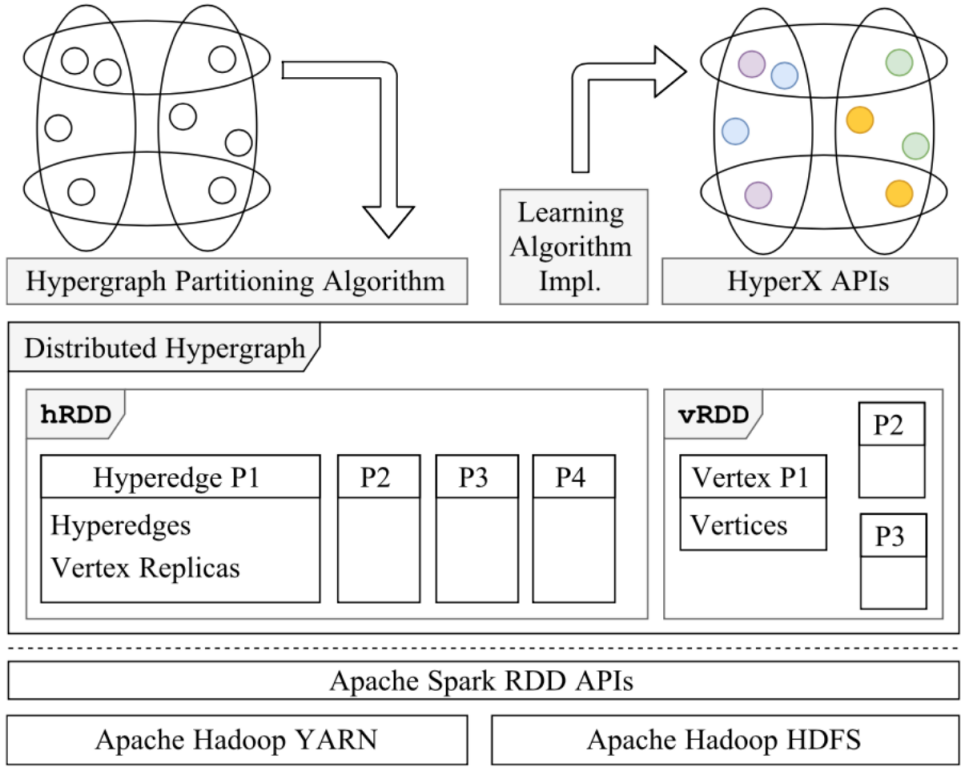


Figure 5.3: Architecture of HyperX

Moreover, HyperX provides a redeployment of Graphx Pregel to develop hypergraph learning algorithms, which introduces a hypergrams program (*hprog*) in addition to the vertices program (*vprog*), already present in the original framework. Hyperedges and vertices have associated values, indicated respectively as *h.val* and *v.val*, where *h.val*

[1] RDD represents a collection of partitioned data elements that can be operated on in parallel

represents the weight of h and *v:val* represents the weight of v. Both *hprog* and *vprog* are executed independently on each partition of the hyperedges and vertices, respectively. When *hprog* and *vprog* are sequentially executed in an iteration, both *h:val* and *v:val* are updated. Finally, HyperX optimizes hyperedges and vertices by partitioning to minimize replicas and to balance workloads. The goal is to minimize memory space and communication costs. Therefore, it uses a new partitioning algorithm based on propagation that is both efficient and effective.

Finally, the *SIMONA* layer handles hypergraph data model of MSN and provides a set of APIs that allows to develop SNA applications (i.e. Community detection, Influence analysis, lurker detection etc.) leveraging the features of the proposed MSN model. This layer allows to combine the social component and the analysis of multimedia content; in fact, it handles the interaction establishes between users and the multimedia content and it leverages the relationship between multimedia objects based on the analysis of high and low-level features. From a technological point of view, this layer is built upon Apache Spark and HyperX to properly manage the entities and relationships involved in the MSN model through an hypergraph data structure.

In particular, this layer extends the capabilities of Hyperx to store both nodes and hyperedges of MSNs as Abstract Data Type (ADT) for handling the different attributes customizable for the specific type of application. For data processing, *Simona* layer combines an iterative vertex-centric approach based on message passing, that allows to process large distributed graphs, and optimization approaches based on sparse matrices.

5.2 Experimental Protocol

In this section, it is described the experimental protocol to provide an evaluation made on real Multimedia Social Network. In particular, the evaluation is made by two parts: first, the feasibility of the proposed model has been proved and the efficacy of the algorithms developed on it and second, the efficacy of the community detection approach has been analyzed.

5.2.1 Dataset

The dataset used for the evaluation is the *Yahoo Flickr Creative Commons 100 Million* (YFCC100m) [138], containing metadata of around 99.2 million photos and 0.8 million videos, shared on Flickr under one of the various *Creative Commons licenses*. Each media object is represented by the following metadata: its Flickr identifier, the user that created it, the camera that took it, the time at which it was taken and when it was uploaded, the location where it was taken (if available), tags and the CC license it was published under.

The Flickr API [2] have been leveraged for gathering users' social information and actions (tags, comments, favorites).

In particular, the examined dataset is composed by the following entity and relationships:

- *Photo Publishing*: author ID, photo ID, Timestamp;
- *Comment Publishing*: author ID, comment ID, photo ID, Timestamp;
- *Favorites*: author ID, photo ID, Timestamp;
- *Photos:* the photo files related to the previous data;

[2]https://www.flickr.com/services/api

- *Similarity Media:* the similarity score between photos.

The *similarities* between images have been computed using the LIRE Image Processing Library (*Lucene Image REtrieval*[3]), an open source visual information retrieval library that allows to compute a *Lucene* index of images' features for Content Based Image Retrieval (CBIR) using local and global state-of-the-art algorithms. In this way, it's possible to compute the similarity of two images as a function: $\delta(Im_1, Im_2) \in [0, 1]$.

The following image describes the possible paths that are considered in our approach.

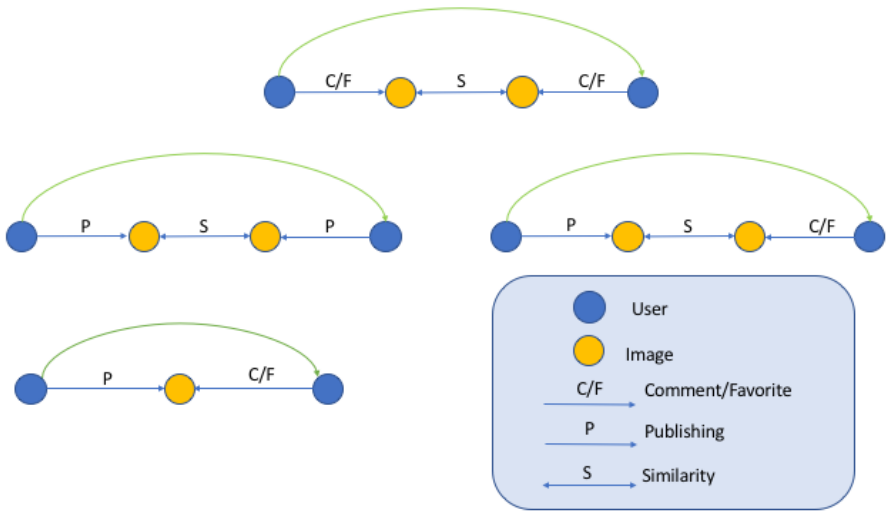


Figure 5.4: Relationships types

In the table 5.1 it is shown the dataset characterization.

[3]http://www.lire-project.net

	Publishing	Comment	Favorites
YFCC100m	42.382.677	11.895.755	8.386.841

Table 5.1: Dataset Characterization

5.2.2 Hardware details

The data crawling process and the image similarities processing have been executed on two engines with Windows 10, Intel i5-6660K with 3.5 GHz (up to 4.5 with turboboost) CPU, Kingstone Hyper-X 16 GB DDR4 memory with 2133 MHz and a NVIDIA GeForce 970 GPU.

The data processing process has been carried out on Microsoft Windows Azure, cloud services provided by Microsoft for the development, management, and hosting of applications and services, using the following configuration: 2 compute optimized instances (F8v2), each one has 8 CPU (2.7 GHz Intel Xeon Platinum 8168 (SkyLake) processor, which can achieve clock speeds as high as 3.7 GHz with the Intel Turbo Boost Technology 2.0) and 16 GB RAM with a 64-bit Ubuntu 14.04 operating system. The proposed technology stack is composed by Hadoop 2.7.3, using YARN as resource manager, and Spark 2.1.11.

5.3 Influence spread Evaluation

To evaluate the feasibility of the proposed model the spread of influence in MSN is analyzed varying seed set cardinality (k) and using TIM+ and IMM, two well-known IM algorithm.

For both algorithms the accuracy value (ϵ) is equal to 0.1. Each experiment has been carried out varying k from 1 to 500 considering the mean over 25 times. This evaluation is carried out in two steps: firstly, the interactions between two users are based on only the well-known

actions (publishing, comment, favorites etc.) and successively also the similarities between the multimedia objects are considered.

The other influence diffusion models that have been compared with our proposal are:

- *Trivalency Model* (TM), that chooses the influence propagation with a random pick inside a set of three values: $p_{u,v} \in [0.1, 0.01, 0.001]$,
- *Weighted Cascade* (WC), where the influence propagation is set to be the in-degree of the node: $p_{u,v} = \frac{1}{|N^{in}(v)|}$.
- *Linear Model*, that corresponds to a linear combination of *reactivity* and *Shareability* as the following:

$$\tau = (\alpha * Reactivity_i + \beta * Shareability_j) = (\alpha * (\frac{n^R_{i,j}}{\sum_{i \in N} n^R_i}) + \beta * (\frac{n^R_{i,j}}{n^A_j})) \tag{5.1}$$

- *Weighted linear model*: that is defined as following:

$$\tau = Norm(Reactivity_i + Shareability_j) = Norm((\frac{n^R_{(i,j)}}{\sum_{i \in N} n^R_i}) + (\frac{n^R_{i,j}}{n^A_j})) \tag{5.2}$$

  To avoid that this parameter assumes a value greater than 1 a normalization technique has been done at the end of the process. One method could be the *Normalization* or the *percentile one*:

$$\tau_i = \frac{\tau_i - \tau_{2,5}}{\tau_{97,5} - \tau_{2,5}} \tag{5.3}$$

  This kind of factor cuts off all the elements before the 2,5 quartile and the ones after the 97,5 quartile.

The spread over the k seed sets over a time window of 2 years using TIM+ and IMM is shown in the following figures 5.5, 5.6.

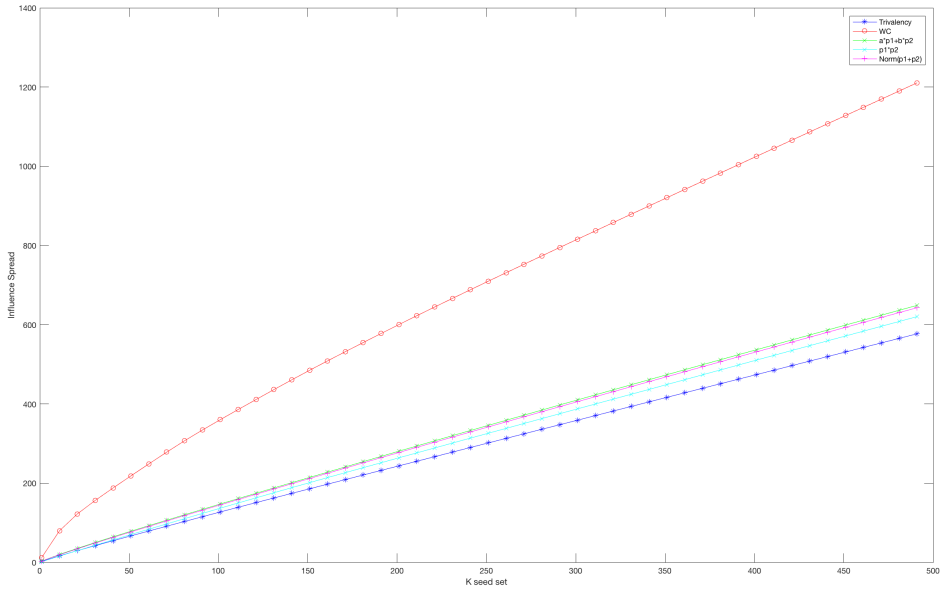


Figure 5.5: TIM: Expected Spread (Time Window 2 Years) varying k Seed Set

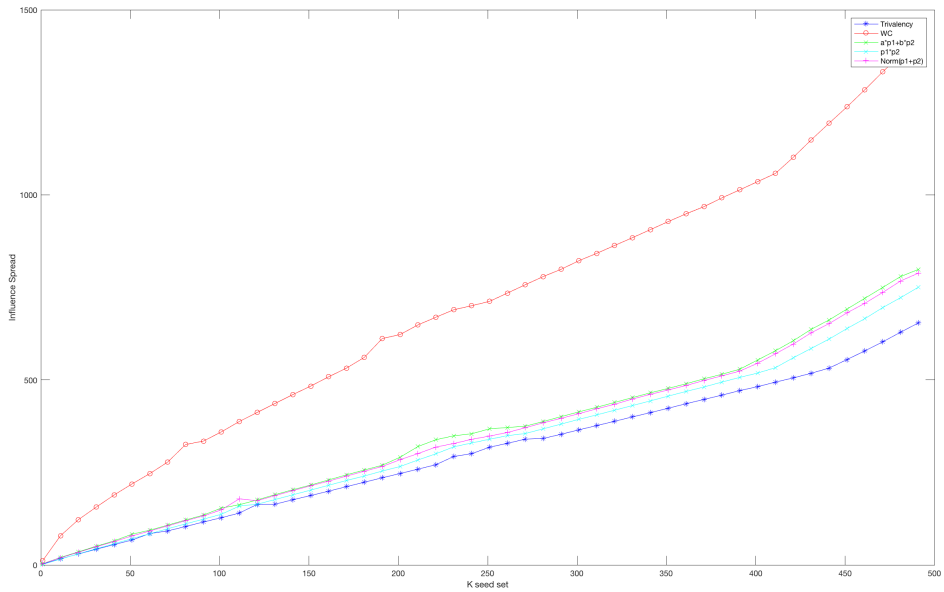


Figure 5.6: IMM: Expected Spread (Time Window 2 Years) varying k Seed Set

In the WC model the propagation probabilities on edges, assuming an high value due to the low value of the nodes in-degree, leading to an high value of expected influence. The influenced spread values of the WC do not reflect the social behavior of users in MSN, because the WC provides too large influence probabilities even if two users interact between them few times.

The influence probabilities used in the proposed model are more similar to the TM because the influence probabilities tend to be small, as shown in [58]. The proposed approach allows to handle the network changes respect to the TM approach, in which the influence values are randomly chosen.

Successively, *similarities* between images have been considered as relationships in the proposed model. If a user v reacts on an image *im1*

that is similar to another picture *im2* posted by another user u, there is a similarity relationship between them.

The following figures 5.7 and 5.8 show the influence spread achieved with these probabilities. Given the increasing of the number of reactions due to the similarities, the difference between the proposed model with respect to the WC decreases, while the difference with the TM increases because the proposed model represents better on social interactions of users, while the WC only considers the topology of the network. Moreover, the proposed model distinguishes with respect to the TM, because it considers the changes inside the network.

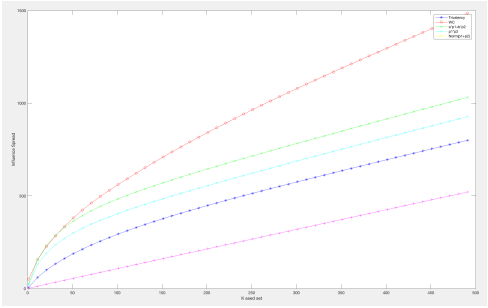


Figure 5.7: TIM: Expected Spread (Time Window 2 Years) varying k Seed Set

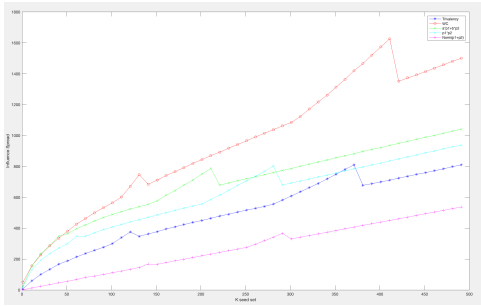


Figure 5.8: IMM: Expected Spread (Time Window 2 Years) varying k Seed Set

As it is easy to note, the propagation probabilities for *linear model* and the proposed approach are increased while the *Weighted linear model* assumes very small values because the normalization factor is affected by outliers leading to have very small influence spread.

5.4 Influence maximization Evaluation

5.4.1 IM bio-inspired approach

To evaluate the effectiveness of the ABC algorithm, several centrality measures for the initial step have been used with respect to a ground truth, computed by ranking Flickr users according to their average number of visualizations, comments and favorites.

Kendall's Tau (τ) and Spearman's Rank Correlation (ρ) coefficients are used to compare the analyzed approaches with respect to the ground truth. As it is possible to note in Table 5.2, the out-degree and degree centrality measures have values more similar to the ground truth but the number of iterations that they required to converge to the final solution is higher respect to other measures.

	τ	ρ
OABC- GR	0,78	0,85
DABC- GR	0,71	0,79
CABC - GR	0,58	0,65
BABC - GR	0,55	0,61

Table 5.2: Ranking comparison: GR(Ground truth ranking), OABC (Out-degree centrality based ABC), DABC (Degree centrality based ABC), CABC (Closeness centrality based ABC), BABC (Betweenness centrality based ABC).

5.4.2 IM approach based on game theory

The evaluation has been performed with increasing complexity of datasets for comparing the different operating modes of the proposed algorithm:

- *Pure Exploration*: This strategy explores the network to discover the efficiency of several arms. In each round the super-arm is chosen randomly, without use information from previous rounds;
- *Pure Exploitation*: This strategy performs exploitation in each round. Only in the first part is possible to perform this action caused by initial lack of knowledge. The super-arms are selected according to the estimated probabilities, through the execution of the *TIM+* method with (1-$\frac{1}{e} - \epsilon$)-approximation guarantee, where $\epsilon = 0.2$;
- ϵ-*Greedy*: The pure exploitation led an high spread triggering a large set of arms. However, after several steps, it is possible that the diffusion stabilizes because a part of the network may remain unexplored. This strategy allows to obtain a trade-off between two strategy above. In short, it combines exploration with probability p_ϵ and exploitation with probability $1 - p_\epsilon$.

Since that the proposed algorithm works on datasets lacking knowledge, the influence probability estimates can be initialized to 0 or to some

prior information. It is chosen $p_{min} \in [0.0010.01]$ as initial estimates of influence probability to allow a faster convergence. It corresponds to a minimum value that can detect when a Trivalency Model (TM) or Weighted Cascade Model (WC) are used to assign the probabilities.

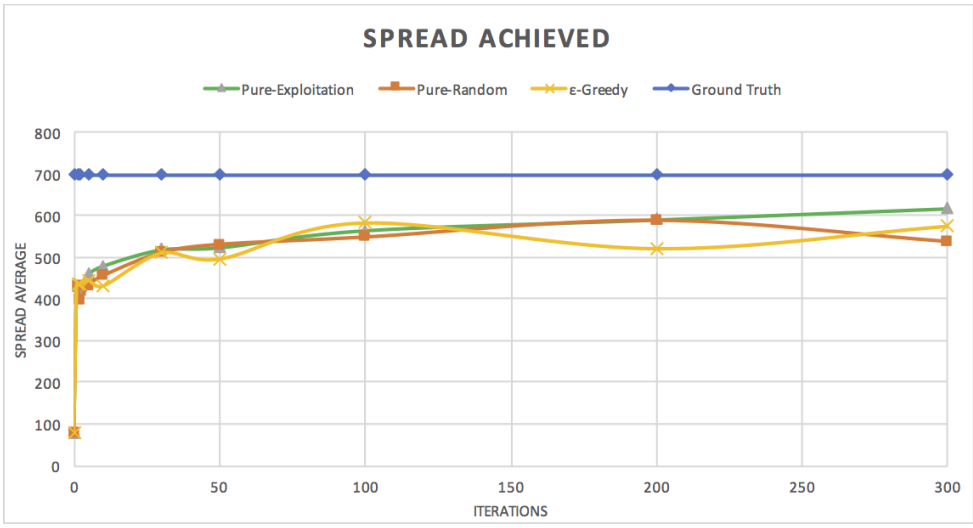
The influence probabilities tend to be small in practice [58], therefore, in the experiments it is set $p_{max} = 0.2$. Moreover, $T = 300$ is chosen as maximum number of rounds and it is used $k = 50$ as cardinality of the seed set. For $\epsilon - Greedy$ strategy it is chosen a $p_o = 25$ that allows to ensure an exploration phase initially.

The proposed approach needs to estimate the spread at each round, after a super-arm is played. For this reason a novel scalable approach based on GraphX Pregel to simulate the Independent Cascade diffusion model. The feedback mechanism plays an important role to update the mean estimate μ and to improve the network knowledge.

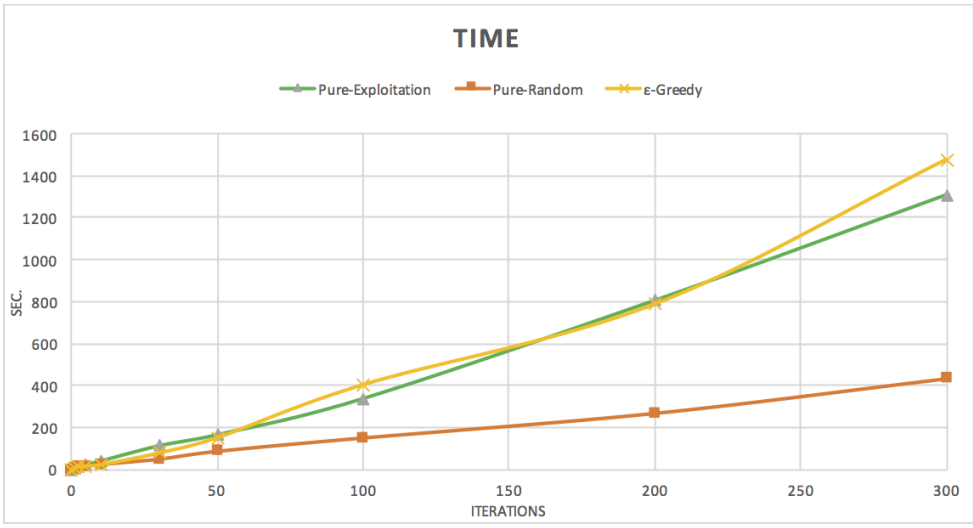
A comparison between the spread obtained with different operating modes of the CMAB, and the "true" spread performing the TIM+ approach. To define the Ground Truth is used *TIM+* algorithm with $\epsilon = 0.1$ and the results are computed by averaging across 25 runs.

In Figure 5.9a it is showed the spread achieved by performing different CMAB modes. The spread increases quickly with the number of rounds and after 300 iterations the bandits approaches become comparable to the Ground Truth (TIM+). In particular the Pure-Exploitation has achieved the best average spread, although it has a slow start with respect to ϵ-Greedy. This latter uses a trade-off between Pure-Exploration and Pure-Exploitation that involves an oscillatory trend. This happens because the explore strategy tries to minimize the local minimum issue. Pure-Exploration has better execution time but the spread increases until to a value, then decreases.

For providing a temporal analysis, in Figure 5.9b it is shown the running time for the various CMAB modes. As it is possible to note, the



(a) *Flickr (WCM) Spread*



(b) *Flickr (WCM) Time*

Figure 5.9: Spread and Time vs Number of rounds

trend for Pure-Exploration is almost linear, the times are expressed in seconds.

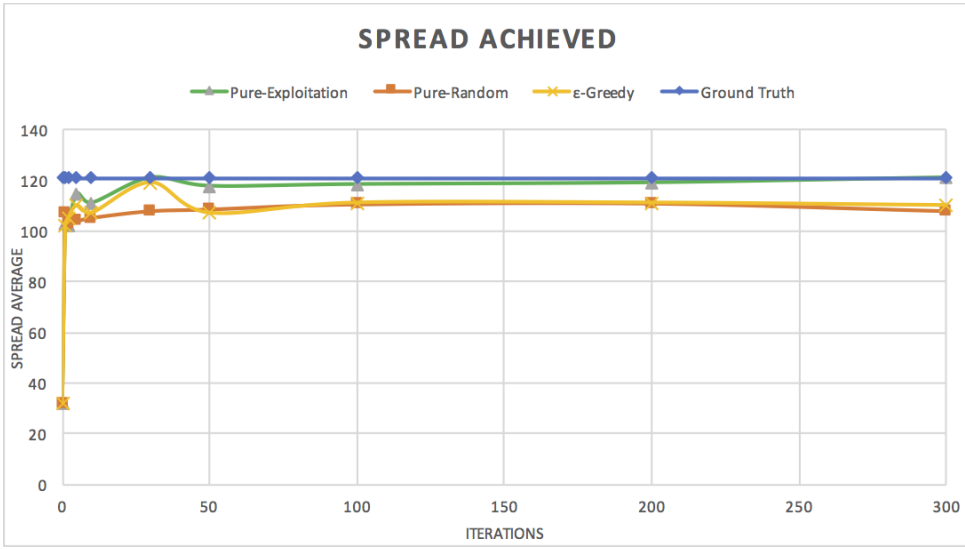
In Figure 5.10, in which the true probability are assigned using the Trivalency Mode, it is possible to observe the same trend but a low spread consequently. Thus, the spread achieved from Pure-Exploitation equals the Ground Truth with only 300 iterations.

5.5 Community Detection Evaluation

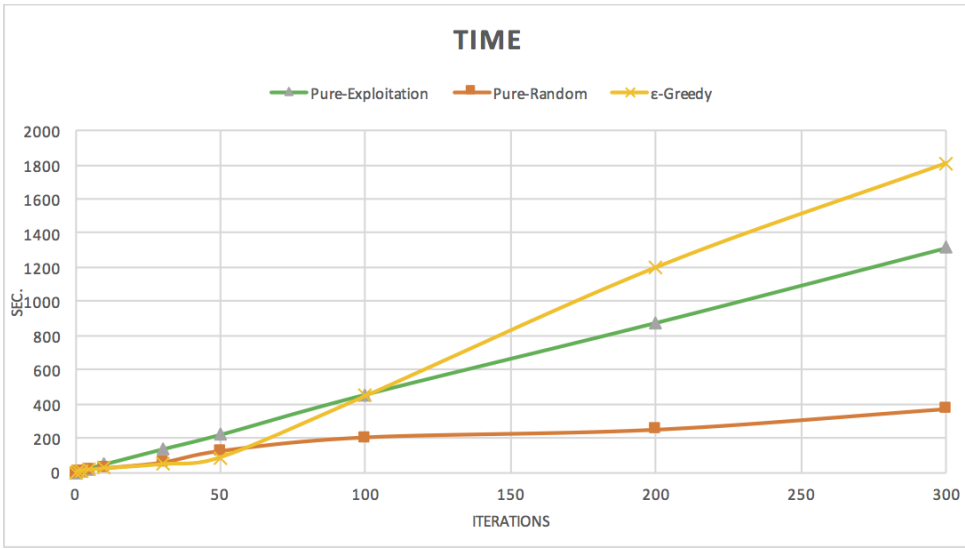
Community Detection algorithms focus on homogeneous networks, while the proposed algorithm leverages the features of MSN model. For this reason, two types of evaluations have been performed:

- Test on homogeneous graph: for the evaluation of the effectiveness of the proposed algorithm. The Zackary dataset[4], a well-known social network of a university karate club composed by 34 members and 78 pairwise, is used to compare the proposed approach with the following algorithms: Infomap, Fast Greedy, Label Propagation and WalkTrap. The quality metrics [45] used for this benchmark are: (i) Normalized Mutual Information (ii) Adjusted Rand Index.
- Test on heterogeneous multimedia social network: this evaluation has been carried out on Flickr dataset. Several communities have been considered based on a given similarity threshold:
  - considering all the relationships except to similarity hyperedges;
  - including similarity hyperedges with weight $p \in [0.90, 1]$;
  - including similarity hyperedges with weight $p \in [0.80, 1]$;
  - including similarity hyperedges with weight $p \in [0.70, 1]$

[4]https://networkdata.ics.uci.edu/data.php?id=105



(a) *Flickr 2(TM) Spread*



(b) *Flickr 2(TM) Time*

Figure 5.10: Spread and Time vs Number of rounds

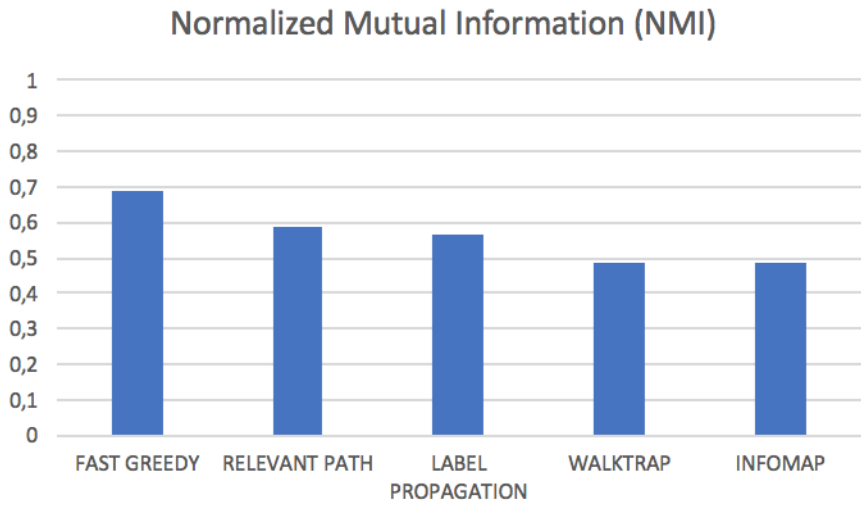


Figure 5.11: Normalized Mutual Information

The figure 5.11 shows the results of Normalized Mutual Information by comparing the well-known community detection algorithms (Fast Greedy, Infomap, Label Propagation and WalkTrap. Our algorith is better than Label Propagation, Walktrap and Infomap) with the proposed approach on the Zackary ground truth.

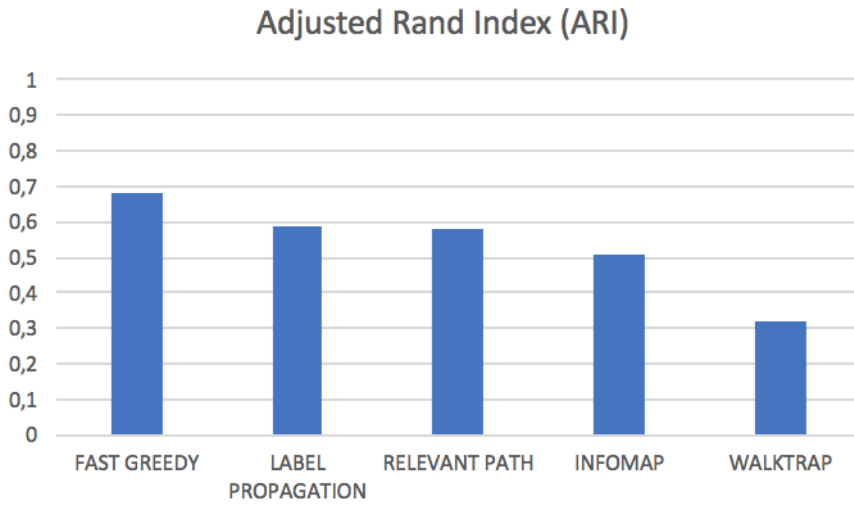


Figure 5.12: Adjusted Rand Index

The result of Adjusted Rand Index are shown in the figure 5.12.

	Proposed Algorithm		**Infomap**		**Fast Greedy**		**Label Propagation**		**Walktrap**	
	TP	**FP**	**TP**	**FP**	**TP**	**FP**	**TP**	**FP**	**TP**	**FP**
Community 1	88,24%	17,65%	94,12%	11,76%	100,00%	5,88%	94,12%	5,88%	52,94%	0,00%
Community 2	82,35%	0,00%	52,94%	5,88%	64,71%	0,00%	64,71%	5,88%	52,94%	0,00%
Community 3	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%	0,00%	100,00%
Community 4									0,00%	100,00%
Community 5									0,00%	100,00%

Table 5.3: Community detection outcome on Zackary Club dataset.

The table 5.3 shows the precision for each community detection algorithm based on the percentage of true positive (TP) and false positive (FP). These values are compared with Zackary ground truth. It is possible to note that the proposed algorithm has the highest percentage about community 2 and a good percentage (88%) about community 1.

The experiment results showed the effectiveness of our algorithm. Instead, the evaluation based on Flickr dataset are shown in table 5.4.

Dataset FLICKR	Number of Communities	Number of hyperedges
Without Similarity	115	9585
Similarity ≥ 0.90	11	13452
Similarity ≥ 0.80	26	29700
Similarity ≥ 0.70	106	224412

Table 5.4: Community Detection on Flickr Dataset

The qualitative analysis on Flickr dataset has been provided in the table 5.4, that shows how the communities number changes when the similarity range increases. The number of communities raises in number when Similarity threshold assumes values greater than 0.7 because the over-fitting of the proposed approach is verified.

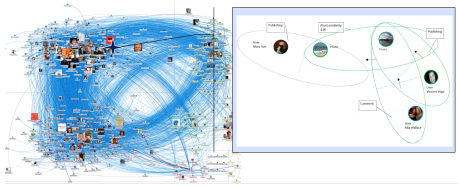


Figure 5.13: Qualitative analysis on Flickr's dataset

The figure 5.13 shows the detection community on sample's dataset. In this sample, two communities have been produced without similarity. Instead, the proposed approach allows to identify one community leveraging similarity relationships.

6 Related Works

In the related works about two main challenges in OSNs are provided and discussed. In particular, we have focused on OSN *Influence Analysis* and *Community detection* research field.

6.1 Influence spread

Influence diffusion analysis describes how the information spreads out across the users of a networks. As described in 2.4.1, different models have been proposed in the literature to model the influence spread in MSNs, that is possible to classify in the following six groups.

1. The first group, relies on the *stochastic process* models, usually based on the study of the probabilistic effect of marketing actions on the initial activation of nodes. The main models are a) the *Threshold* model [59], in which a node is activated only if the sum of influence provided by its neighbor is greater than specific node threshold, and b) the *Independent Cascade* model [57], in which a node has only one chance to activate its neighbors.

2. The second group is composed by *Epidemic model*, in which the influence spread is designed as a disease spread among biological populations by a given *infection rate*. The approaches are based on epidemic dynamics definition, taking into account the entire population. A first attempt to model influence diffusion using epidemic spreading is the Susceptible Infected-Susceptible model, *SIS* [4], that has only two possible transitions: the first one $S \rightarrow I$ occurs when a susceptible individual interacts with an infectious individual and becomes infected and the second $I \rightarrow$ occurs when the infectious individual recovers from the disease and returns to the pool of susceptible individuals. These two transitions could involved an individual multiple time. Another approach is Susceptible infected-recovered, *SIR*, model. In this approach when an individual recovers from disease it acquires a permanent immunity or is removed. In both models the infection process involves a contact among two or more users while the others processes occur spontaneously after a certain time. Differently, the Susceptible-Infected-Removed-Susceptible or *SIRS* model, provides a temporary immunity.

3. The *Voter Model*[129] forms the third family, in which each node can switch its state based on the opinions of its neighbors. In [87] the authors propose a voter model on signed networks to analyze analyze the dynamics of influence diffusion of two opposite opinions. the proposed approaches exploit voter model to analyze the spread diffusion on the network. A first attempt to analyze the influence diffusion using signed network is provided in [86], where the authors exploit this type of network to analyze the spread of two opposite opinions. Moreover, Even-Dar et al. [47] propose some algorithms in the contest of probabilistic voter model.

4. The fourth group is formed by the *Markov Random Field* models, in which each node v has a binary variable X_v that depends on its neighbors N_v and is independent on the non-neighbors nodes.

5. Moreover, the influence spread can be also modeled by *heat models*[42], in which the influence spread is described as a heat diffusion process by a diffusion probability assigned to each node.

6. Eventually, the influence diffusion problem can be analyzed as a *bond percolation* [66, 69], in which the edges are designed as bonds and nodes as sites. In this group there are two types of approaches: bond percolation, that perform the evaluation taking in account that an edge could be either open or closed independently, and site percolation, in which sites can be opened or closed independently. In particular, the bond percolation theory can be seen as an *Independent Cascade* (IC) model where the influence diffusion can be viewed as bond percolation from the seed set. Thus, these models consider graphs where edges or nodes can be independently open or closed with certain probabilities [21]. The authors [110] present the exact solution of the problem for the specific, but highly relevant, case of the Susceptible-Infected-Removed (SIR) model for epidemic spreading at criticality. By exploiting the mapping between bond percolation and the static properties of SIR, the authors prove that the Non-Backtracking centrality is the optimal criterion for the identification of influential spreaders in locally tree-like networks at criticality. Finally, the authors [97] map the problem onto optimal percolation in random networks to identify the minimal set of influencer, which arises by minimizing the energy of a many-body system, where the form of the interactions is fixed by the non-backtracking matrix of the network.

Following a data-driven view of OSNs, we have considered and analyzed mainly the *"user-to-content"* relationships that describe interactions between users and the content they generate. The proposed solution, described in 4.1.1, is a novel diffusion algorithm that has a first phase in which it tries to learn from the network the likelihood that users can influence each others. It is an *Action-Reaction Based* solution, that analyzes the action made by two users on same or similar multimedia objects to estimate the likelihood that a user can influence another one. The aim of the approach is to represent the process in which a given user changes its behavior based on the information disseminated from its peers; for instances, a user that buys or adopts a new product may influence the behavior of its friends. The proposed approach belongs to the category of stochastic models because the influence that one user exerts on another one is based on stochastic analysis of the users' actions made on the MSN. In particular, our approach is more similar to the *Independent Cascade* model (IC) given that the diffusion process is based on the τ values that represent how interact two users.

6.2 Influence Maximization

Deciding whether to adopt an innovation (such as a political idea or product), individuals are frequently influenced, explicitly or implicitly, by their social contacts. Indeed, the way in which new practices spread through a population depends mainly on the fact that people influence each others behavior. It is essential for companies to target "opinion leaders", as influencing them will lead to a large cascade of further recommendations. This is the goal of each viral marketing and social advertisement campaigns, and corresponds in solving the *influence maximization problem*. The Influence Maximization is the problem of finding

a small subset of nodes that could maximize the spread of influence over the network.

The influence maximization approaches can be classified in the following groups: *Stochastic*, *Bio-Inspired*, *Game Theory*, *Genetic Algorithm* and *Community Detection*

The *Stochastic* approaches are based on the stochastic diffusion models. The well-known algorithms are the following:

- TIM+ [137], a two-phase influence maximization algorithm based on a *sketch* method. In the first phase the θ parameter is derived using a lower bound respect to the maximum expected spread; in the second phase the computed θ value is exploited to derive a k-size set of nodes that covers a large number of Reverse Random (RR) sets.

- IMM[135]: an algorithm leveraging the two-phases approach of TIM that exploits a set of estimation probabilistic techniques [151] to compute a lower bound of the maximum expected influence of any size-k set of nodes that is asymptotically tight.

The algorithm 2 based on the proposed influence model belongs to this category of IM approaches because it uses the RR sets to compute the k influencers. In particular, the novelties of the approach is to reduce the amount of required Reverse Random (RR) sets based on the possible small set of relevance of multimedia objects shared in MSN, computed as a function of the interaction on them between users.

The second category of influence maximization algorithms is composed by bio-inspired approaches based on collective behavior of social animals, that could interact among them and with their environment, such as bee, ant and so on. Yang et al.[158] propose a first approach based on *ant colony optimization* algorithm to deal with Influence Maximization problem. Moreover, Zhang et al. [160] provide an algorithm

based on Particle Swarm Optimization (PSO) exploiting social interaction pattern to find the most influential users in micro-blogging services. Eventually, a method that simulates the bee behavior for simulate of influence propagation in viral campaigns through micro-blogging platform is proposed in [121]. In particular this approach combines the global-local search capacity of the ABC algorithm with a node ranking procedure.

The third family is based on the game theory approaches for oligopolistic market or bandit theory. In [162] the authors propose an approach, taking in account jointly viral product attributes and viral influence attributes, based on Stackelberg Game Theory for product adoption maximization and product line performance optimization. Variants of influence maximization bandit have also been studied by [81], [144]. The first approach uses a different objective of maximizing the expected size of the union of the influenced nodes over time, while the second one discusses node level feedback rather than the edge level feedback. However, as for regret minimization, their results are mostly empirical and there is no theoretical regret bounds for us to compare with. Another approach based on game theory is proposed in [104], where the authors deal with two problems about social influence analysis. Firstly, they propose a near-optimal polynomial-time seeding algorithms for three representative classes of social network models with the aim to identify a good subset of individuals to seed for reducing the diffusion time significantly. Successively, they propose a practical seeding algorithm, called Practical Partitioning and Seeding (PrPaS), for reducing the diffusion time can be reduced by choosing a good set of users. Bao et al.[15] propose a *RSB* algorithm based on multi-armed bandit optimization to maximize influence propagation over time, calibrating its explore-exploit strategy utilizing outcomes of previous decisions. Another approach based on bandit theory is provided in [148], in which the analyzed problem combines

challenges of limited feedback, due to limited vision of network by agent, and combinatorial number of actions, due to the cardinality of feasible set is exponential in the maximum number of influencers. To deal with this problem, Wen et al. [149] propose a IMLinUCB algorithm with a regret bound polynomial in all quantities, reflecting the structure of the network and the probabilities of influence An evolution of combinatorial multi-armed bandit with probabilistically triggered arms (CMAB-T) and semi-bandit feedback is proposed by [145] to overcome the issue presented in the prior CMAB-T studies where the regret bounds contain a possibly exponentially large factor of $1/p^*$, where p^* is the minimum positive probability that an arm is triggered by any action.

Another family of approaches is based on Genetic algorithm. In particular, Bucur et al. [24] show that, by using simple genetic operators, it is possible to find in feasible runtime solutions of high-influence that are comparable, and occasionally better, than the solutions found by a number of known heuristics (one of which was previously proven to have the best possible approximation guarantee, in polynomial time, of the optimal solution). The advantages of Genetic Algorithms show, however, in them not requiring any assumptions about the graph underlying the network, and in them obtaining more diverse sets of feasible solutions than current heuristics. In [150], the authors propose evolutionary algorithms, implemented on GPGPU, for selecting seeds in social networks, showing that it is possible to outperform well-known greedy algorithm in the problem of influence maximization for linear threshold model in both: quality (up to 16% better) and efficiency (up to 35 times faster). Lu et al.[92] develop an algorithm, that recursively estimates the influence spread using reachable probabilities from node to node, to enable greedy algorithms to perform well in big social network influence maximization, providing strategies that integrate memory cost and computing efficiency.

The last family is based on community detection algorithm for reducing the solution space in which identify the seed set. Rahimkhani et al. [112] propose an algorithm, called *ComPath*, based on the linear threshold model with the aim to reduce the number of investigated nodes exploiting the structural communities of the underlying network. ComPath+[12] improves the previous algorithm reducing the number of nodes analyzed and preserving the quality of seed. To address the problem of performance guarantee, Shang et al.[127] propose *CoFIM* a community-based framework for influence maximization on large-scale networks. The influence propagation process is divided into two phases (i) seeds expansion, that analyze the expansion of seed nodes among different communities at the beginning of diffusion; and (ii) intra-community propagation, that analyze the influence propagation within communities which are independent of each other.

Our proposed diffusion model allows to use any category of models described above. In particular, in the algorithm 4 it is provided a modified ABC algorithm based on the proposed influence model considering that bees can exert their influence by several ways, such as publishing a content, posting a review, etc.. Moreover, some changes have been introduced respect to the original algorithm: the k most influent nodes are computed by assuming that a waggle dance was successfully performed by a user u_i on a user u_j based on the action that they made on the same or similar multimedia objects. Furthermore, to solve the local maximum issue, a *tabu search* technique is used in each iteration; in particular, if the number of scouts is not equal to a given user defined parameter, the algorithm chooses a random number of scouts. In addition, several centrality measures can be used for the ranking stage.

Finally, an approach based on bandit theory is proposed in section 4.1.2, that allows to deal with the problems where no influence probabilities are known. The proposed approach overcomes the lack of knowledge

using only the graph structure, in which the use of τ values allows to obtain a faster convergence of the proposed algorithm. This approach allows to combine the learning of influence with spread optimization.

6.3 Community Detection

Nowadays, community mining techniques have been successfully applied to multimedia-related applications, such as user modeling, photo tagging, video annotation, recommendation, targeted advertising etc. In particular, the analysis of user interactions and user community can significantly improve performances of several applications, such as On-line recommendation of friends and multimedia objects. The community detection algorithms focus on homogeneous social networks, i.e., each network node, representing a multimedia resource or a user while in the multimedia social network everything changes because there are multi-typed objects and relationships. For this reason, in section 4.2.1, We propose a community detection algorithm belonging to the community quality optimization class that identifies communities in a heterogeneous network (MSN).

As described in chapter 2.4.2, there is no single, universally accepted definition of community within a social network. The ability to detect the community structure of networks plays an important role in the analysis of complex systems; in particular, identifying communities provides more information about how the network is organized but it is almost impossible to know or even to estimate the number of communities in a Social Media network.

The first approaches were based on graph partitioning (e.g., Metis [73], Graclus [40]), that divides the vertices set in n disjoint groups following some optimization metrics (i.e. minimize the number of edges lying between the groups (*cut size*)). The proposed approach is different for

two fundamental aspects: (i) it do not require the number of communities as input, but it defines the number of communities automatically as outputs; (ii) the identified groups could be overlapped, that is pairwise intersections do not result always in the empty set.

Other approaches attempt to identify vertices groups that are more densely connected to each other than to the leftovers nodes of the network. The biggest family of community detection algorithms is formed by those based on maximizing modularity [101], that corresponds to identify communities with an internal edge density larger than that expected in a given graph model. Several strategies have been proposed for its optimization, such as agglomerative greedy [31] or simulated annealing [93]. The approach on modularity optimization scales to graphs with hundreds of millions of objects, as shown in [20], but its results decrease considerably as long as the size of the graph increases [79] and it is not able to detect small and well defined communities when the graph is too large. Finally, it has been reported that modularity has resolution limits [13] [49].

Other approaches rely on Random walks techniques; the idea is that the probability of remaining into a community is higher than going outside due to the higher density of internal edges. Walktrap [107] and Infomap [119] belong to this category of community detection approaches. In particular, the latter is one of the best community detection algorithm that leverages a specific code for describing random walks based on communities is searched. The codification that requires less memory space (attains the highest compression rates) is selected. These approaches are mainly focused on the topology of the network while the proposed technique considers also the semantic information related to nodes through new relationships produced by combining high-level and low-level features.

However, in many social and information networks, nodes can belong to multiple communities. Thus, another category is composed by algorithms able to find overlapping communities. An example of such an algorithm is Oslom [80], which uses the significance, defined as the probability of finding a given cluster in a random null model, as a fitness measure to assess the quality of a cluster. Another algorithm that falls into this category is Link Clustering Algorithm (LCA) [2], that is based on the idea of taking edges instead of vertices to identify a community. The overlapped communities are built by an iterative process in which the similarity of adjacent edges (i.e. those edges that share a vertex, forming an opened triad) is assessed by using the Jaccard coefficient of the adjacency lists of the two other vertices of the edges. Moreover Label propagation [111] is another family of iterative techniques that initially sets labels to nodes. Successively, it defines rules that simulate the spread of these labels in the network similarly to infections. Finally, an algorithm, called *BigClam* is proposed by Yang et al. [155]. This algorithm is based on computing an affiliation of vertices to communities that maximizes an objective function using non negative matrix factorization. The objective function is based on the intuition that the probability of existing of an edge between two vertices increases with the number of communities the vertices share (i.e. the number of communities in which the vertices overlap).

Natively, our approach does not handle overlapping communities but they can be integrated into the proposed approach by partitioning multimedia contents within the analyzed network.

7 Conclusions

This dissertation was mainly focused on the design of novel data model relying on hypergraph data structure to support in a simple way all the different kinds of relationships that are typical of Multimedia Social Network. This model provides a solution for representing MSN sufficiently general with respect to: i) a particular social information network, ii) the different kinds of entities, iii) the different types of relationships, iv) the different applications (such as Viral Marketing, recommendation systems, community detections and so on).

In particular the features of the proposed model have been used to deal with the following two challenges: *Influence Maximization* and *Community detection*. Concerning the first problem, a novel influence diffusion model has been proposed that, learning recurrent user behaviors from past logs, estimates the probability that a user can influence the other ones, basically exploiting user to content actions. The main contribution about the IM problem is related to the definition of a novel influence diffusion model, whose behavior is similar on the one hand to TM, given that the exerted influence from one user respect to another one tends to be small because that it is mediated by the actions made by it on the network, and on other hand it is similar to WC,

because the proposed approach considers the network topology handling also semantic information about actions made by users. Moreover, the proposed approach focuses on users behaviors by defining of an influence operator that attempt to integrate sociology and psychology aspects which affect human behaviors in the proposed model. Exploiting the features of this model, several IM algorithms (based on game theory, epidemiological etc.) have been developed.

The feasibility of the model and effectiveness of the IM proposed approaches have been evaluated on *Flickr*, a real Multimedia Social Network, using our Big Data platform based on Spark technology. This evaluation shows how the proposed approaches can be properly faced IM problem leveraging the introduced model.

Regarding the second challenge, the most of the community detection algorithms focus on homogeneous social network while the proposed algorithm leverages both user interactions and multimedia content, in terms of high and low-level features, for identifying the hidden communities in Multimedia Social Network (MSN). The main contributions about community detection problem are related to:

- A novel model of Multimedia social networks that integrates both information about users of one or more OSNs and the content related to it generated and shared by a hypergraph data structure.
- A community detection algorithm based on heterogeneous network, that leverages also similarity relationships between two multimedia objects to provide a new way to build the Weighted User Matrix used by the examined algorithm.

To evaluate the effectiveness of the approaches two types of evaluations have been performed: the first used for evaluate the effectiveness with respect to well-known community detection algorithm; the second one carried out to provide a qualitative analysis of the proposed algorithm on Multimedia Social Network varying the similarity threshold.

The provided evaluation shows on one hand how our approach has similar performance with respect to the well-known algorithms in literature while on other hand it analyzed how the proposed approach allows to handle heterogeneous network and the interaction between different entities of a MSN and the advantages to introduce similarity relationships based on the choose of an appropriate similarity threshold.

Bibliography

[1] Adalı, S., Lu, X., and Magdon-Ismail, M. (2014). Local, community and global centrality methods for analyzing networks. *Social Network Analysis and Mining*, 4(1):1–18.

[2] Ahn, Y.-Y., Bagrow, J. P., and Lehmann, S. (2009). Link communities reveal multiscale complexity in networks. *arXiv preprint arXiv:0903.3178*.

[3] Albors, J., Ramos, J. C., and Hervas, J. L. (2008). New learning network paradigms: Communities of objectives, crowdsourcing, wikis and open source. *International Journal of Information Management*, 28(3):194–202.

[4] Allen, L. J. (1994). Some discrete-time si, sir, and sis epidemic models. *Mathematical biosciences*, 124(1):83–105.

[5] Amato, F., Mazzeo, A., Moscato, V., and Picariello, A. (2013). A framework for semantic interoperability over the cloud. In *Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on*, pages 1259–1264. IEEE.

[6] Amato, F., Mazzeo, A., Penta, A., and Picariello, A. (2008). Using nlp and ontologies for notary document management systems. In *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on*, pages 67–71. IEEE.

[7] Amato, F., Moscato, V., Picariello, A., Piccialli, F., and Sperlí, G. (2017a). Centrality in heterogeneous social networks for lurkers detection: An approach based on hypergraphs. *Concurrency and Computation: Practice and Experience*.

[8] Amato, F., Moscato, V., Picariello, A., and Sperlí, G. (2016). Multimedia social network modeling: A proposal. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, pages 448–453.

[9] Amato, F., Moscato, V., Picariello, A., and Sperlı, G. (2017b). Diffusion algorithms in multimedia social networks: a preliminary model.

[10] Andersen, R. and Lang, K. J. (2006). Communities from seed sets. In *Proceedings of the 15th international conference on World Wide Web*, pages 223–232. ACM.

[11] Asendorpf, J. B. and Wilpers, S. (1998). Personality effects on social relationships. *Journal of personality and social psychology*, 74(6):1531.

[12] Bagheri, E., Dastghaibyfard, G., and Hamzeh, A. (2016). An efficient and fast influence maximization algorithm based on community detection. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 1636–1641.

[13] Bagrow, J. P. (2012). Communities and bottlenecks: Trees and treelike networks have high modularity. *Physical Review E*, 85(6):066118.

[14] Bansal, N., Blum, A., Chawla, S., and Meyerson, A. (2004). Approximation algorithms for deadline-tsp and vehicle routing with time-windows. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 166–174. ACM.

[15] Bao, Y., Wang, X., Wang, Z., Wu, C., and Lau, F. C. (2016). Online influence maximization in non-stationary social networks. In *Quality of Service (IWQoS), 2016 IEEE/ACM 24th International Symposium on*, pages 1–6. IEEE.

[16] Barnes, J. A. (1954). Class and committees in a norwegian island parish. *Human relations*, 7(1):39–58.

[17] Baumeister, R. F. and Finkel, E. J. (2010). *Advanced social psychology: The state of the science*. OUP USA.

[18] Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009). Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 49–62. ACM.

[19] Berry, D. A. and Fristedt, B. (1985). *Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)*, volume 12. Springer.

[20] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

[21] Bollobás, B. and Riordan, O. (2006). *Percolation*. Cambridge University Press.

[22] Borgs, C., Brautbar, M., Chayes, J., and Lucier, B. (2014). Maximizing social influence in nearly optimal time. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 946–957. Society for Industrial and Applied Mathematics.

[23] Brocke, J. v., Richter, D., and Riemer, K. (2009). Motives for using social network sites (snss)-an analysis of sns adoption among students. *BLED 2009 Proceedings*, page 40.

[24] Bucur, D. and Iacca, G. (2016). Influence maximization in social networks with genetic algorithms. In *European Conference on the Applications of Evolutionary Computation*, pages 379–392. Springer.

[25] Burke, M., Marlow, C., and Lento, T. (2009). Feed me: motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 945–954. ACM.

[26] Chakraborty, T., Dalmia, A., Mukherjee, A., and Ganguly, N. (2017). Metrics for community analysis: A survey. *ACM Computing Surveys (CSUR)*, 50(4):54.

[27] Chen, W., Wang, Y., and Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM*

SIGKDD international conference on Knowledge discovery and data mining, pages 199–208. ACM.

[28] Chen, W., Wang, Y., Yuan, Y., and Wang, Q. (2016). Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778.

[29] Cialdini, R. B. (2001). Science and practice.

[30] Cialdini, R. B. (2016). *Pre-Suasion: A Revolutionary Way to Influence and Persuade*. "Simon & Schuster".

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

[32] Collins, R. (1988). The micro contribution to macro sociology. *Sociological Theory*, 6(2):242–253.

[33] Cornuejols, G., Fisher, M. L., and Nemhauser, G. L. (1977). Exceptional paper—location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management science*, 23(8):789–810.

[34] Corstjens, M. and Umblijs, A. (2012). The power of evil. *Journal of Advertising Research*, 52(4):433–449.

[35] Coscia, M., Rossetti, G., Giannotti, F., and Pedreschi, D. (2012). Demon: a local-first discovery method for overlapping communities. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 615–623. ACM.

[36] Costa, P. T. and MacCrae, R. R. (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI): Professional manual*. Psychological Assessment Resources, Incorporated.

[37] Costa, P. T. and McCrae, R. R. (1992). Four ways five factors are basic. *Personality and individual differences*, 13(6):653–665.

[38] Costenbader, E. and Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social networks*, 25(4):283–307.

[39] De Choudhury, M., Sundaram, H., John, A., and Seligmann, D. D. (2009). What makes conversations interesting?: themes, participants and consequences of conversations in online social media. In *Proceedings of the 18th international conference on World wide web*, pages 331–340. ACM.

[40] Dhillon, I. S., Guan, Y., and Kulis, B. (2007). Weighted graph cuts without eigenvectors a multilevel approach. *IEEE transactions on pattern analysis and machine intelligence*, 29(11).

[41] Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM.

[42] Doo, M. and Liu, L. (2014). Probabilistic diffusion of social influence with incentives. *IEEE Transactions on Services Computing*, 7(3):387–400.

[43] Dwyer, C., Hiltz, S., and Passerini, K. (2007). Trust and privacy concern within social networking sites: A comparison of facebook and myspace. *AMCIS 2007 proceedings*, page 339.

[44] Ellison, N. B. et al. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

[45] Emmons, S., Kobourov, S., Gallant, M., and Börner, K. (2016). Analysis of network clustering algorithms and cluster quality metrics at scale. *PloS one*, 11(7):e0159161.

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

[47] Even-Dar, E. and Shapira, A. (2007). A note on maximizing the spread of influence in social networks. In *International Workshop on Web and Internet Economics*, pages 281–286. Springer.

[48] Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3):75–174.

[49] Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41.

[50] Fournier, S. and Avery, J. (2011). The uninvited brand. *Business horizons*, 54(3):193–207.

[51] French, J. R. and Raven, B. (2004). The bases of social power. *Studies in social power*, pages 1959–150.

[52] Gai, Y., Krishnamachari, B., and Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking (TON)*, 20(5):1466–1478.

[53] Galaskiewicz, J. and Wasserman, S. (1993). Social network analysis: Concepts, methodology, and directions for the 1990s. *Sociological Methods & Research*, 22(1):3–22.

[54] Garey, M. R. and Graham, R. L. (1974). Performance bounds on the splitting algorithm for binary testing. *Acta Informatica*, 3(4):347–355.

[55] Gilovich T., Keltner D., N. R. (2011). *Social psychology*. W.W. Norton & Company.

[56] Golbeck, J., Robles, C., and Turner, K. (2011). Predicting personality with social media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 253–262, New York, NY, USA. ACM.

[57] Goldenberg, J., Libai, B., and Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223.

[58] Goyal, A., Bonchi, F., and Lakshmanan, L. V. (2010). Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM.

[59] Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443.

[60] Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American journal of sociology*, 91(3):481–510.

[61] Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.

[62] Gulbahce, N. and Lehmann, S. (2008). The art of community detection. *BioEssays*, 30(10):934–938.

[63] Gunelius, S. (2010). *30-Minute Social Media Marketing: Step-by-step Techniques to Spread the Word About Your Business: Social Media Marketing in 30 Minutes a Day*. McGraw Hill Professional.

[64] Gupta, S. and Kumar, P. (2016). Community detection in heterogenous networks using incremental seed expansion. In *Data Science and Engineering (ICDSE), 2016 International Conference on*, pages 1–5. IEEE.

[65] Heimo, T., Kumpula, J. M., Kaski, K., and Saramäki, J. (2008). Detecting modules in dense weighted networks with the potts method. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(08):P08007.

[66] Hu, Y., Havlin, S., and Makse, H. A. (2014). Conditions for viral influence spreading through multiplex correlated social networks. *Physical Review X*, 4(2):021031.

[67] Huang, J., Zhang, R., and Yu, J. X. (2015). Scalable hypergraph learning and processing. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 775–780. IEEE.

[68] Ji, X., Wang, Q., Chen, B.-W., Rho, S., Kuo, C. J., and Dai, Q. (2014). Online distribution and interaction of video data in social multimedia network. *Multimedia Tools and Applications*, pages 1–14.

[69] Jiang, Y. and Jiang, J. (2015). Diffusion in social networks: A multiagent perspective. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(2):198–213.

[70] Jung, K., Heo, W., and Chen, W. (2012). Irie: Scalable and robust influence maximization in social networks. In *2012 IEEE 12th International Conference on Data Mining*, pages 918–923. IEEE.

[71] Kang, C., Molinaro, C., Kraus, S., Shavitt, Y., and Subrahmanian, V. (2012). Diffusion centrality in social networks. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 558–564. IEEE Computer Society.

[72] Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.

[73] Karypis, G. and Kumar, V. (1995). Metis–unstructured graph partitioning and sparse matrix ordering system, version 2.0.

[74] Katz, D. and Kahn, R. L. (1978). *The social psychology of organizations*, volume 2. Wiley New York.

[75] Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. *Public opinion quarterly*, 21(1):61–78.

[76] Kempe, D., Kleinberg, J., and Tardos, É. (2003a). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM.

[77] Kempe, D., Kleinberg, J., and Tardos, E. (2003b). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 137–146, New York, NY, USA. ACM.

[78] Krishnamurthy, B. (2009). A measure of online social networks. In *Communication Systems and Networks and Workshops, 2009. COMSNETS 2009. First International*, pages 1–10. IEEE.

[79] Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117.

[80] Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Finding statistically significant communities in networks. *PloS one*, 6(4):e18961.

[81] Lei, S., Maniu, S., Mo, L., Cheng, R., and Senellart, P. (2015). Online influence maximization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 645–654. ACM.

[82] Leskovec, J., Adamic, L., and Huberman, B. (2007a). The dynamics of viral marketing acm trans. Web.

[83] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007b). 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.

[84] Lewin, K. (1936). Some social-psychological differences between the united states and germany. *Journal of Personality*, 4(4):265–293.

[85] Lewin, K. (1951). Field theory in social science: selected theoretical papers (edited by dorwin cartwright.).

[86] Li, Y., Chen, W., Wang, Y., and Zhang, Z.-L. (2013). Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 657–666. ACM.

[87] Li, Y., Chen, W., Wang, Y., and Zhang, Z.-L. (2015). Voter model on signed social networks. *Internet Mathematics*, 11(2):93–133.

[88] Li, Y.-M., Lai, C.-Y., and Chen, C.-W. (2011). Discovering influencers for marketing in the blogosphere. *Information Sciences*, 181(23):5143–5157.

[89] Litt, E. (2013). Understanding social network site users' privacy tool use. *Computers in Human Behavior*, 29(4):1649–1656.

[90] Liu, D., Ye, G., Chen, C.-T., Yan, S., and Chang, S.-F. (2012). Hybrid social media network. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 659–668. ACM.

[91] Liu, H. (2007). Social network profiles as taste performances. *Journal of Computer-Mediated Communication*, 13(1):252–275.

[92] Lu, W.-X., Zhou, C., and Wu, J. (2016). Big social network influence maximization via recursively estimating influence spread. *Knowledge-Based Systems*, 113:143–154.

[93] Medus, A., Acuña, G., and Dorso, C. (2005). Detection of community structures in networks via global optimization. *Physica A: Statistical Mechanics and its Applications*, 358(2):593–604.

[94] Mitchell, J. C. (1969). *Social networks in urban situations: analyses of personal relationships in Central African towns*. Manchester University Press.

[95] Moradi, F. (2014). *Improving Community Detection Methods for Network Data Analysis*. Chalmers University of Technology.

[96] Moreno, J. L., Jennings, H. H., et al. (1934). *Who shall survive?*, volume 58. JSTOR.

[97] Morone, F. and Makse, H. A. (2015). Influence maximization in complex networks through optimal percolation. *Nature*, 524(7563):65–68.

[98] Moscato, V., Picariello, A., and Subrahmanian, V. (2015). Multimedia social networks for cultural heritage applications: the givas project. In *Data Management in Pervasive Systems*, pages 169–182. Springer.

[99] Nan, G., Zang, C., Dou, R., and Li, M. (2015). Pricing and resource allocation for multimedia social network in cloud environments. *Knowledge-Based Systems*, 88:1 – 11.

[100] Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294.

[101] Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

[102] O'Donovan, F. T., Fournelle, C., Gaffigan, S., Brdiczka, O., Shen, J., Liu, J., and Moore, K. E. (2013). Characterizing user behavior and information propagation on a social multimedia network. In *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, pages 1–6. IEEE.

[103] Ohsaka, N., Akiba, T., Yoshida, Y., and Kawarabayashi, K.-i. (2014). Fast and accurate influence maximization on large networks with pruned monte-carlo simulations. In *AAAI*, pages 138–144.

[104] Ok, J., Jin, Y., Shin, J., and Yi, Y. (2016). On maximizing diffusion speed over social networks with strategic users. *IEEE/ACM Transactions on Networking*, 24(6):3798–3811.

[105] Papadopoulos, S., Kompatsiaris, Y., Vakali, A., and Spyridonos, P. (2012). Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3):515–554.

[106] Pentland, A. (2015). *Social Physics: How social networks can make us smarter*. Penguin.

[107] Pons, P. and Latapy, M. (2006). Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10(2):191–218.

[108] Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Notices of the AMS*, 56(9):1082–1097.

[109] Pujol, J. M., Erramilli, V., Siganos, G., Yang, X., Laoutaris, N., Chhabra, P., and Rodriguez, P. (2010). The little engine (s) that could: scaling online social networks. *ACM SIGCOMM Computer Communication Review*, 40(4):375–386.

[110] Radicchi, F. and Castellano, C. (2016). Leveraging percolation theory to single out influential spreaders in networks. *Physical Review E*, 93(6):062314.

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

[112] Rahimkhani, K., Aleahmad, A., Rahgozar, M., and Moeini, A. (2015). A fast algorithm for finding most influential people based on the linear threshold model. *Expert Systems with Applications*, 42(3):1353–1361.

[Rice] Rice, R. Diffusion of innovations.

[114] Richter, D., Riemer, K., and vom Brocke, J. (2011). Internet social networking. *Wirtschaftsinformatik*, 53(2):89–103.

[115] Rodrigues, T., Benevenuto, F., Cha, M., Gummadi, K., and Almeida, V. (2011). On word-of-mouth based discovery of the web. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 381–396. ACM.

[116] Rogers, E. M. (1983). *Diffusion of innovations*. Free Press ; Collier Macmillan.

[117] Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.

[118] Ross, C., Orr, E. S., Sisic, M., Arseneault, J. M., Simmering, M. G., and Orr, R. R. (2009). Personality and motivations associated with facebook use. *Computers in human behavior*, 25(2):578–586.

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

[120] Rosvall, M. and Bergstrom, C. T. (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, 6(4):e18209.

[121] Sankar, C. P., Asharaf, S., and Kumar, K. S. (2016a). Learning from bees: An approach for influence maximization on viral campaigns. *PloS one*, 11(12):e0168125.

[122] Sankar, C. P., S., A., and Kumar, K. S. (2016b). Learning from bees: An approach for influence maximization on viral campaigns. *PLOS ONE*, 11(12):1–15.

[123] Schneider, F., Feldmann, A., Krishnamurthy, B., and Willinger, W. (2009). Understanding online social network usage from a network perspective. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 35–48. ACM.

[124] Schultz, D. E. and Peltier, J. (2013). Social media's slippery slope: challenges, opportunities and future research directions. *Journal of research in interactive marketing*, 7(2):86–99.

[125] Shamma, D. A., Kennedy, L., and Churchill, E. (2010). Summarizing media through short-messaging services. In *Proceedings of the ACM conference on computer supported cooperative work*, pages 551–552.

[126] Shamma, D. A., Shaw, R., Shafton, P. L., and Liu, Y. (2007). Watch what i watch: using community activity to understand content. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 275–284. ACM.

[127] Shang, J., Zhou, S., Li, X., Liu, L., and Wu, H. (2017). Cofim: A community-based framework for influence maximization on large-scale networks. *Knowledge-Based Systems*, 117:88–100.

[128] Snow, D. A., Zurcher Jr, L. A., and Ekland-Olson, S. (1980). Social networks and social movements: A microstructural approach to differential recruitment. *American sociological review*, pages 787–801.

[129] Sood, V. and Redner, S. (2005). Voter model on heterogeneous graphs. *Physical review letters*, 94(17):178701.

[130] Spagnoletti, P., Resca, A., and Sæbø, Ø. (2015). Design for social media engagement: insights from elderly care assistance. *The Journal of Strategic Information Systems*, 24(2):128–145.

[131] Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

[132] Tajfel, H. (1982). Social psychology of intergroup relations. *Annual review of psychology*, 33(1):1–39.

[133] Tang, Y., Shi, Y., and Xiao, X. (2015a). Influence maximization in near-linear time: A martingale approach. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1539–1554. ACM.

[134] Tang, Y., Shi, Y., and Xiao, X. (2015b). Influence maximization in near-linear time: A martingale approach. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 1539–1554, New York, NY, USA. ACM.

[135] Tang, Y., Shi, Y., and Xiao, X. (2015c). Influence maximization in near-linear time: A martingale approach. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 1539–1554, New York, NY, USA. ACM.

[136] Tang, Y., Xiao, X., and Shi, Y. (2014a). Influence maximization: Near-optimal time complexity meets practical efficiency. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 75–86. ACM.

[137] Tang, Y., Xiao, X., and Shi, Y. (2014b). Influence maximization: Near-optimal time complexity meets practical efficiency. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 75–86, New York, NY, USA. ACM.

[138] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. (2016). Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

[139] Tian, Y., Srivastava, J., Huang, T., and Contractor, N. (2010). Social multimedia computing. *Computer*, 43(8):27–36.

[140] Tsvetovat, M. and Kouznetsov, A. (2011). *Social Network Analysis for Startups: Finding connections on the social web*. "O'Reilly Media, Inc.".

[141] Tuten, T. L. (2008). *Advertising 2.0: Social Media Marketing in a Web 2.0 World: Social Media Marketing in a Web 2.0 World*. ABC-CLIO.

[142] Urban, G. (2005). *Don't just relate-advocate!: a blueprint for profit in the era of customer power*. Pearson Education.

[143] Van Dongen, S. M. (2001). *Graph clustering by flow simulation*. PhD thesis.

[144] Vaswani, S., Lakshmanan, L., Schmidt, M., et al. (2015). Influence maximization with bandits. *arXiv preprint arXiv:1503.00024*.

[145] Wang, Q. and Chen, W. (2017). Tighter regret bounds for influence maximization and other combinatorial semi-bandits with probabilistically triggered arms. *arXiv preprint arXiv:1703.01610*.

[146] Wang, Y., Cong, G., Song, G., and Xie, K. (2010). Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1039–1048. ACM.

[147] Weimann, G. (1994). *The influentials: People who influence people*. SUNY Press.

[148] Wen, Z., Kveton, B., and Valko, M. (2016a). Influence maximization with semi-bandit feedback. *arXiv preprint arXiv:1605.06593*.

[149] Wen, Z., Kveton, B., and Valko, M. (2016b). Influence maximization with semi-bandit feedback. *CoRR*, abs/1605.06593.

[150] Weskida, M. and Michalski, R. (2016). Evolutionary algorithm for seed selection in social influence process. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1189–1196.

[151] Williams, D. (1991). *Probability with martingales*. Cambridge university press.

[152] Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P., and Zhao, B. Y. (2009). User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. Acm.

[153] Xie, J. and Szymanski, B. K. (2012). Towards linear time overlapping community detection in social networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 25–36. Springer.

[154] Yang, B., Liu, D., and Liu, J. (2010). Discovering communities from social networks: Methodologies and applications. In *Handbook of social network technologies and applications*, pages 331–346. Springer.

[155] Yang, J. and Leskovec, J. (2013). 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.

[156] Yang, J. and Leskovec, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213.

[157] Yang, Q., Chen, W. N., Yu, Z., Gu, T., Li, Y., Zhang, H., and Zhang, J. (2017). Adaptive multimodal continuous ant colony optimization. *IEEE Transactions on Evolutionary Computation*, 21(2):191–205.

[158] Yang, W.-S., Weng, S.-X., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2012). Application of the ant colony optimization algorithm to the influence-maximization problem. *International Journal of Swarm Intelligence and Evolutionary Computation*, 1(1).

[159] Yu, B., Chen, M., and Kwok, L. (2011). Toward predicting popularity of social marketing messages. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 317–324. Springer.

[160] Zhang, B., Zhong, S., Wen, K., Li, R., and Gu, X. (2013). Finding high-influence microblog users with an improved pso algorithm. *International Journal of Modelling, Identification and Control*, 18(4):349–356.

[161] Zhang, Z. and Wang, K. (2013). A trust model for multimedia social networks. *Social Network Analysis and Mining*, 3(4):969–979.

[162] Zhou, F., Jiao, R. J., and Lei, B. (2016). Bilevel game-theoretic optimization for product adoption maximization incorporating social network effects. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(8):1047–1060.

List of figures

List of tables