

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
UNIVERSITY MOHAMED BOUDIAF - M'SILA
FACULTY of Mathematics and Computer Science
DEPARTMENT: Computer Science
BRANCH: Computer Science
OPTION: SIGL



Dissertation submitted to obtain Master degree

By: Miss. ZEBBICHE Lydia.

SUBJECT

**The analysis of information diffusion in
social media networks:
A comparative experimental study**

Supported before the jury composed of:

MR. LOUCIF Hemza	University of M'sila	Supervisor
MR. BENAZI Makhlouf	University of M'sila	Examiner
MR. DEBBA Ali	University of M'sila	Examiner

Academic year: 2020/2021

Dedication:

I would like to dedicate this humble thesis to my mother. the person who always gave me hope and believe in me.

“Look mom, I did make it “

Acknowledgement:

All throughout writing of this thesis, I got a great lot of help and support.

Firstly, I would like to pay my special regards to my supportive supervisor, Dr. Loucif Hemza for providing guidance and feedback. He gave me the opportunity to learn so much about the domain of the analysis of social networks. Without his persistent help, the goal of this thesis would not have been realized.

I also wish to acknowledge the support and great love of my family, my mother Theldja, my father, Mohamed El Hadi, My sisters, Lynda and Khitem, and my dearest brothers Mounir, Mourad, Omar and Nabil. They kept me going on and this work would not have been possible without their input.

Finally, I would like to thank my colleagues and friends, especially Hadjer, Sofiane and Loubna who provided fascinating chats as well as pleasant distractions to take my attention away from my research.

Table of contents:

Dedication:	2
Acknowledgement:	3
Liste of tables:	6
Liste of figures:	7
General Introduction:	9
Chaptre I. Introduction to Social Media Networks:	10
1. Introduction:	10
2. What's social media?	10
3. Social Media VS Social Networks	11
4. History[4]	11
5. Types of Social Media:	13
6. Social Media Platforms Functions:[1]	13
7. Summary:	14
Chaptre II. Introduction to SNA and its background:	16
1. Introduction:	16
2. Definition of social network analysis:	16
3. Social Network Analysis Applications:	17
4. Graph theory fundamentals:[7]	17
5. Representing and Visualising networks data:	20
6. Types of network:[6]	23
7. Social networking analysis metrics:	25
8. Summary:	28
Chaptre III. Information Diffusion Models:	30
1. Introduction:	30
2. Information diffusion in social networks:	30
3. Diffusion models in social networks:[26] [17]	32
a) Independent Cascade Model (ICM):	33
b) Linear Threshold Model (LTM)[31]:	33
c) Linear Influence Model (LIM)[32]:	34
d) Epidemic Model:	36
4. Summary:	38
Chaptre IV. Implementation:	40
1. Introduction:	40
2. Information cascades:	40

3. Modelling information diffusion:	41
4. How to choose initial active set:	44
5. Experiments tools:	50
6. Datasets:	54
7. Summary:	56
Chapitre V. Experiment's Results:	58
1. Introduction:	58
2. Network's structure and metrics analysis:	58
3. Summary:	63
General conclusion:	65
Bibliography:	66
Abstract:	70
Resume:	Error! Bookmark not defined.
ملخص:	72

Liste of tables:

Table 1.1: Adjacency list example.

Table 4.1: The structure characteristics of Facebook users.

Table 4.2: The centrality metrics.

Table 4.3: The nodes activated before 1000 runs (average).

Liste of figures:

Figure 1.1: Two graphs (a) an undirected graph, and (b) a directed graph.

Figure 1.2: A weighted graph.

Figure 1.3: Bridges and structural hole [5].

Figure 1.4: Graph and its corresponding Adjacency matrix.

Figure 1.5: The edge list corresponding to the graph exhibited by figure 1.4.

Figure 1.6: Full and egocentric network [16].

Figure 1.7: Multimodal networks.

Figure 1.8: Examples of computing the density of a network.

Figure 1.9: Closeness centrality of a node.

Figure 1.10: Betweenness centrality of a node.

Figure 2.1: An overview of information diffusion models [17].

Figure 2.2: An illustration of the Linear Influence Model (LIM).

Figure 2.3: SIR model showing the three phases.

Figure 2.4: Illustration of the SIR Epidemic model [26].

Figure 3.1: Linear Threshold Model example [38].

Figure 3.2: Degree appearance on my graph.

Figure 3.3: In-degree appearance on my graph.

Figure 3.4: Out-degree appearance on my graph.

Figure 3.5: List of high degree that I used it in my experiment.

Figure 3.6: List of high betweenness that I used it in my experiment.

Figure 3.7: List of randomly choose.

Figure 3.8: The complete Gephi workspace window.

Figure 3.9: LTM by Python.

Figure 3.10: ICM by Python.

Figure 3.11: Graph of my datasets.

Figure 3.12: Subgraphs of different degrees.

Figure 3.13: The edges list shows relationships among the nodes.

Figure 4.1: Facebook user's centrality metrics distribution plots.

Figure 4.2: The results of betweenness.

Figure 4.3: The results of degree.

General Introduction:

Online social networks have dramatically increased in recent years. Which is described as the method users interact with each other. These interactions include: an individual creating content and sharing it with its friends, re-sharing the contents shared by an individual's friends, submitting comments or liking friends' postings.

This research focuses on information propagation models for anticipating number of infected nodes in a network and knowing how information spreads according to different models (comparative study) and which model is the quickly easily one on propagating the information, where we've used two modes Independent Cascade Model (ICM) The Linear Threshold Model (LTM),

This thesis is organized in five chapters. For each of them a brief description is written as follows:

Chapter One

We present social media networks in a comprehensive way.

Chapter Two

Introduces some significant definitions in social network analysis domain starting by graph theory essentials for understanding network's structure. It also carries the network representations and visualization in addition to networks types and SNA major metrics.

Chapter Three

Consecrated about information diffusion and influence definitions and models.

Chapter Four

Devote about implementation it carries information diffusion modelling, the tools (Gephi, Python), the used datasets (Facebook).

Chapter Five

Show the experiments steps and analysis results and the comparison of the two models.

Finally, a **General Conclusion** recapitulates a brief summary of the thesis.

Chapitre I. Introduction to Social Media Networks:

1. Introduction:

Due to the current demographic boost, human life is consequently founded on social relationships which form a network of individuals, cultures, and organizations.

The social interaction and communication experience has greatly expanded as well as the web and the hard evolution that change the early social network based on face-to-face interactions to a less difficult online social network based on information and communication technologies tools.

2. What's social media?

Social Media is a web-based technology that uses online tools in order to facilitate the sharing of content, profiles, experiences, thoughts, knowledge and media itself. Therefore, it eases conversations and online interaction between most people who interconnect in different ways on social media platforms nowadays. In **social media** platform we include:[1]

- **Facebook**
- **Instagram**
- **Twitter**
- **Linked in**
- **Pinterest**
- **Reddit**

and many more platforms are available there.

Social media can be seen as a valuable source of knowledge, thoughts, and attitudes in a variety of fields of study. For instance, in May 2019,

more than 500 hours of video were uploaded to YouTube every minute. This equates to approximately 30,000 hours of newly uploaded content per hour¹. There are 500 million tweets per day and around 200 billion tweets per year ². There are much more online platforms than humans on the planet. In addition, technical and economic frameworks are now based on social media, which has made it extremely difficult to predict and reason for their behaviours.

3. Social Media VS Social Networks

Social Media can be called a strategy and an outlet for broadcasting, while Social Networking is a tool and a utility for connecting with others.

Cohen (2009) reports that "the difference is not just semantics but in the features and functions put into these websites by their creators which dictates the way they are to be used." [2]

In fact, there are several differences between social media and social networks .The first one could be the definition; social media is still a media which is primarily used to transmit or share information with a broad audience, while social networking is an act of engagement as people with common interests associate together and build relationships through community .[3]

4. History[4]

Social media was further developed during the 1970s.[3]

1971: Email

There are ongoing debates on whether Email could be considered as a part of social media. We cannot ignore the fact that the introduction of Email marked the beginning to the much more collaborative social media years later.

¹ <https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/>

² <https://www.dsayce.com/social-media/tweets-day/>

1979: Usenet

Usenet is a worldwide distributed Internet discussion system. It developed from the general-purpose UUCP³ architecture of the same name.

1984: LISTSERV

LISTSERV scaled up the usage of email communication and proved as an effective way to reach out to a large number of people in an instant.

1988: IRC

Email, Usenet and Listserv helped in sharing messages, individually, categorized and also to a group through a list. This still could not replicate the ease of meeting a person face to face. Internet Relay Chat (IRC) introduced in 1988 improved the experience. IRC is a form of chatting. Originally designed for group chatting in discussion forums, this also allowed one-to-one communication via private message as well as chat and data transfers.

1991: Personal websites, Discussion groups, chat

After the success of IRC, many personal websites, discussion groups and chat groups also became popular. People were thrilled at the possibility of sharing their opinions and often went overboard in their expressions. The first online social media etiquette standards were proposed, and called netiquette, as a control mechanism.

1995: Social Networking site classmate.com

The objective of this social media website is to assist members in finding friends and acquaintances from throughout their lives, from kindergarten, primary school, high school, college, work and the United States military. Classmates.com has over 50 million members.

1998-2004 Blogs, Podcast, Wikis and Facebook[5]

The modern blog evolved from the online diary. Mark Zuckerberg founded in 2004 Facebook which allows its users to enter personal information

³ Unix-to-Unix Copy

and interact with other users.[5] According to recent statistics, approximately 2,234 billion⁴ online users are interested on this fascinating platform monthly.

2005 and beyond

Major launches⁵ : YouTube (2005), Twitter (2006), Tumbler (2007), Pinterest& Instagram (2010), Snapchat (2011) and Vine (2013).

5. Types of Social Media:

There are several distinctive types of social media[6]:

Threaded conversations (emails, Usenet newsgroups), WWW (homepages), cooperative authoring (Wiki, sharing documents), blogs and podcasts (LiveJournal, Twitter), social sharing (YouTube, Flickr, Last.fm, reddit), social networking services (Facebook, LinkedIn), online markets and production (eBay, Amazon, Top coder), virtual worlds (The Sims, Second life, World of Warcraft), idea generation, mobile based services (location sharing and games) ...

6. Social Media Platforms Functions:[1]

Social media networks are widely used for communication and knowledge exchange, but times have changed. In today's world, social media platforms are used not just for networking but also for a variety of other purposes, including:

- ✓ Use for Business growth.
- ✓ Use for advertising and marketing of product and services.
- ✓ Use for E-commerce.
- ✓ Use of E-learning.
- ✓ Use to improve client relationships.
- ✓ Use for link building.

⁴ <https://www.blogdumoderateur.com/50-chiffres-medias-sociaux-2019/>

⁵ <https://scalar.usc.edu/works/everything-you-always-wanted-to-know-about-social-media-but-were-too-afraid-to-ask/history-and-different-types-of-social-media>

7. Summary:

The presence of social networks in the social life of human being cannot be ignored. Thanks to the internet and the new web technologies, they turned from being simply a tool of communication into a way of exploring and exchanging great issues in different domains.

In the coming chapter, we intend to deal with SNA and its background in a detailed way.

Chaptre II. Introduction to SNA and its background:

1. Introduction:

As we have seen in the precedent chapter, an introduction to social media networks and their progress in several domain and contexts. and to get a deeper understanding of the data that flows through those social networks, data analysis is needed.

Throughout this section, we would like to emphasize that the purpose of this chapter is to illustrate the fundamentals of SNA, which will assist us in interpreting social network architectures based on graph theory principles and social network analysis measures.

we will focus on graph theory concepts in order to learn about the structure of social networks. we start with SNA and its applications. Then we discuss network representation and visualization, and its different possible types. Following that, we list the main and most critical social network research metrics. The aim of this chapter is to understand the basics of SNA to be able to interpret social network structures based on graph theory concepts and social network analysis measures.

2. Definition of social network analysis:

Social network analysis (related to network theory) has emerged as a key technique in modern sociology. It has also gained a significant following in anthropology, biology, communication studies, economics, geography, information science, organizational studies, social psychology, and sociolinguistics, and has become a popular topic of speculation and study[7].

Social network analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities[8].

3. Social Network Analysis Applications:

Social network analysis methods have been extended to a pretty wide variety of domains, and they have become an extremely common personal and business involvement. SNA is being used in business applications to help manage business challenges, gain insight into markets and communities, and create tighter industry relationships[9]. According to Patti Anklam, network maps offer a detailed snapshot of a business environment at a certain moment in time, which aids in answering several questions in the community-building process, such as: Is the network able to make decisions quickly and easily? Is the network continuously creating leadership capacity? Which businesses will provide a better return on investment? Who are playing leadership roles in the community? Who is not, but should be?[10].

Another significant application of SNA is measuring influence, diffusion and contagion. Contagions that flow through human-based networks can be bad (disease, gossip), good (ideas and information) or neutral (money and investments).

In [11], they have selected a subset of the extremely large number of SNA applications domains, such as : Health Care , Marketing , Tourism and Hospitality , Cyber Security , Emerging Area(Politics , Detection of fake news and misinformation , and Multimedia)...

4. Graph theory fundamentals:[7]

In [12] it's mentioned that : "Network analysis, which is based on graph theory and matrices, is an enormously versatile way of understanding social systems underpinned by a rich variety of techniques".

A graph is a way of specifying relationships among a collection of items. It consists of a set of objects, called *nodes*, with certain pairs of these objects connected by links called *edges*[13]. For example, the graph in Figure 1.1 shows two simple graphs. In social networks, vertices (also named nodes, agents, entities or actors) often represent people or social structures such as workgroups, teams, organizations, states or even countries. The links that connect vertices (also called edges, ties, relationships or simply connections) are the building blocks of networks. In a social network, there are two main types of links: strong links that correspond to friends, called strong ties, and weak links that correspond to acquaintances, called weak ties.

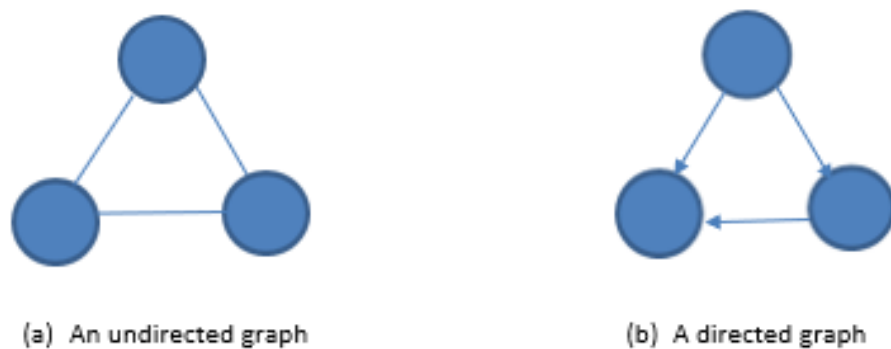


Figure 1.1: Two graphs: (a) an undirected graph, and (b) a directed graph.

Directed and Undirected graphs: a directed graph is a graph where its nodes are connected by directed edges (arrows) that represent asymmetric relationships. In undirected graph, the edges represent symmetric relationships, with the direction is not important. Figure (1.1) (a) illustrates an undirected graph, while figure (1.1) (b) shows a directed graph.

A path is essentially a list of nodes and the edges that connect them, with the property that each pair of nodes in the sequence is linked by an edge. A path that does not repeat nodes is referred to as a simple path.

Length of the path is the number of moves it has from beginning to finish, referring to the number of edges in the sequence that makes it up. The length of a path is used to determine if two nodes in a graph are close or far apart (or a network). The distance is the shortest path between two nodes.

A diameter refers to the longest shortest path between any pair of nodes in this graph. In some other manners, diameter is the length of longest distance of the graph.

A walk is similar to a path, except that each node can be visited many times.

A cycle is a path with at least three edges, in which the first and the last nodes are the same. Otherwise, it is a walk such that the origin and destination are the same.

A network: a network can be defined as a graph in which nodes and/or edges have attributes (e.g. names)⁶.

Connected graph: When a path exists between two distinct nodes, the graph is said to be connected. Or else, it's defined as "disconnected."

Weighted graph is a graph in which each edge is associated with a numerical value called a weight (see Figure (1.2))

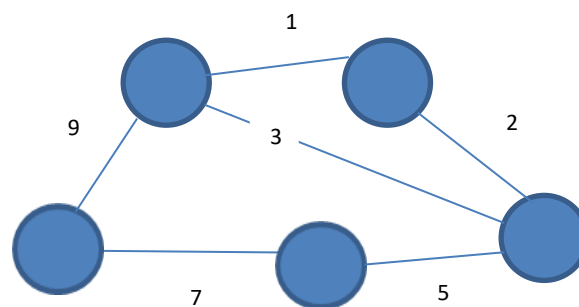


Figure 1.2: A weighted graph.

Bridges: an edge is said to be a bridge if deleting it would cause its endpoints to lie in different components of a graph.

⁶ https://en.wikipedia.org/wiki/Network_theory

Structural hole: in a network, we call a node a structural hole if it is connected to multiple local bridges. In Figure (1.3), the node A is with her multiple local bridges.

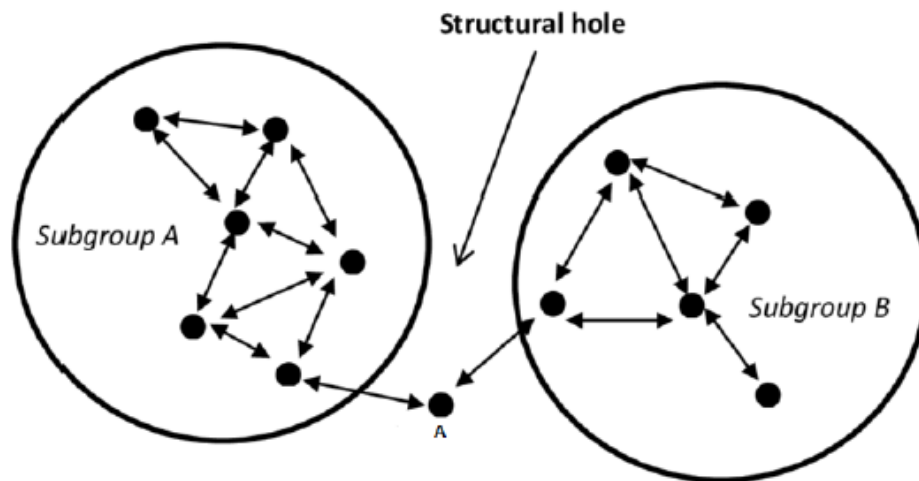


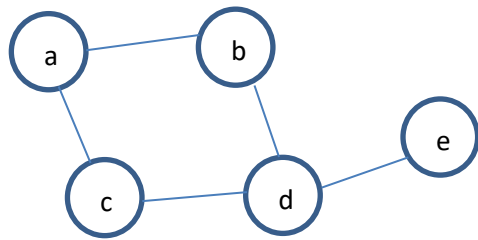
Figure 1.3: Bridges and structural hole[5]

5. Representing and Visualising networks data:

Graphs are useful ways of presenting information about social networks because they serve as mathematical models of network structures[13].

Visual representation of social networks is important to understand the network data and convey the result of the analysis[14]. Representing the information also allows the application of mathematical and computer tools to summarize and find patterns.

- ✓ **Adjacency Matrix:** is a squared matrix $N \times N$ array in which rows and columns represent the same sets of nodes. If there is no relationship between vertex v_i and v_j then the value is 0. If a relationship exists between two vertices, then the value is either 1 in the case of non-weighted graphs, or is the weight of the link between the pair of vertices in the case of a weighted graph. For an undirected graph, the adjacency matrix is symmetric.



Undirected Graph

	a	b	c	d	e
a	0	1	1	0	0
b	1	0	0	1	0
c	1	0	0	1	0
d	0	1	1	0	1
e	0	0	0	1	0

Adjacency Matrix

Figure 1.4: A Graph and Its Corresponding Adjacency Matrix.

- ✓ **Adjacency List:** each vertex has a list of which vertices it is adjacent to. This causes redundancy in an undirected graph: for example, if vertices A and B are adjacent, A's adjacency list contains B, while B's list contains A. Adjacency queries are faster, at the cost of extra storage space[7]. For the undirected graph shown above the corresponding adjacency list is shown in table 1.

Nodes	Connected to
a	b , c
b	a , d
c	a, d
d	b, c, e
e	d

Table 1: Adjacency List example.

- ✓ **Edge List:** Is another common approach for storing graphs as an array of two numbers of the nodes that the edges of the graph are incident on. For sparse graphs, both the adjacency and edge lists are the suitable choice for their representation. For the graph shown in Figure 1.4., we have the following edge list representation:

(a,b)
(a,c)
(b,a)
(b,d)
(c,a)
(c,d)
(d,b)
(d,c)
(d,e)
(e,d)

Figure 1.5: The Edge List corresponding to the graph exhibited by Figure 1.4.

Visualisation is extremely beneficial to social network researchers in terms of comprehending new ways to present and manage data, as well as efficiently converting data into useful knowledge[15].

- ✗ Networks can be visualized in many different ways.
- ✗ There are a variety of tools available for visualizing networks.
- ✗ Visualizations can help provide a fast and global understanding of the network.
- ✗ It is important to be careful when interpreting and evaluating visualizations because they typically do not tell the whole story about networks.

6. Types of network:[6]

- **Egocentric networks:**

In a network, the individual who is the centre of attention is called “ego”, and the people he or she is connected to “alters.” Egocentric networks are a type of network that includes people who are linked to the ego. Such as, your personal Facebook network because you are connected to all your friends.

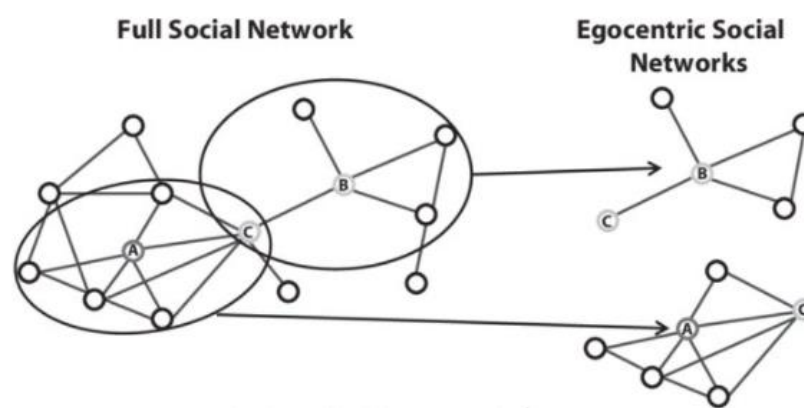


Figure 1.6: Full and egocentric networks[16].

- **Full networks:**

It contains and links between all persons or institutions of interest, where all egos are similarly treated. It is also called complete network and normally created and available when a single system acts as a hub among a group of connected people.

For example, the whole population of the planet could be considered as a single social network, although it would be difficult to capture all these connections[15].

- **Unimodal networks:**

are the standard networks that include one type (mode) of nodes. They connect users to users or they connect documents to documents, but they don't include both users and documents.

Figure (1.7) represents Unimodal networks, where nodes are represented by the same colour to describe that they are of the same type.

- **Multimodal networks:**

are the networks that connect different type of nodes. Such as, connecting a user to a blog, or a person to an organization. In this type of networks, it is preferable to use various shapes and colours to separate different types of nodes. A common type of multimodal network is a bimodal network with exactly two types of nodes.

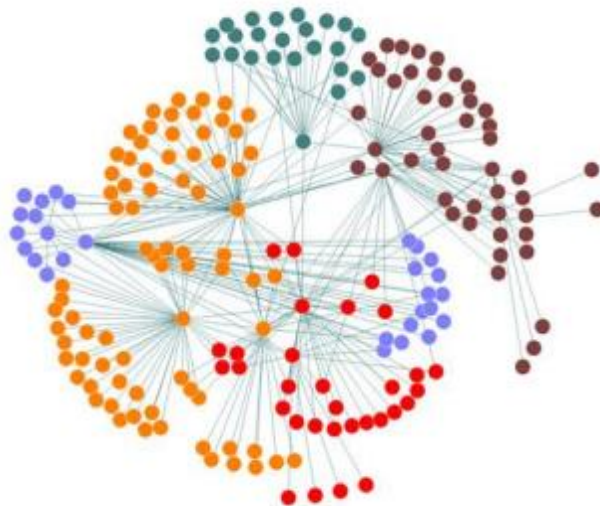


Figure 1.7: Multimodal networks.

- **Affiliation networks:**

Affiliation network is bimodal network that include individuals and some event, activity, or content with which they are affiliated. For example, an affiliation network may link users to the wiki pages they have edited. Pages include individuals affiliated with them. This kind of network can be transformed to user-to-user network,

and affiliation-to-affiliation network (article-to-article in the example of wiki pages). Generally, this approach can be used to relate objects of all types.

- **Multiplex networks:**

they are networks that include many types of connections. Good example in this type is Twitter network because that can contain three different types of directed edges: 'following' relationships, 'reply to' relationships, and 'mention' relationships.

7. Social networking analysis metrics:

Social network Analysis is the research area which has emerged as a key technique that focuses on measuring and mapping relationships and information flows within the networks that connect people or social units (users, groups, organizations, teams, computers and other entities) to a formal models, besides the study of those units and their attributes [6][17]. Experts in the field of social networks have worked to develop theories and algorithms for measuring new quantitative social network metrics. These metrics help to understand the global structure of the network.

- **Density:** is a measure of the connectedness and (the level of interconnectedness of vertices) in a network. Density is a count of the number of relationships observed to be present in a network divided by the total number of possible relationships that could be presented [18]. In the figure (1.8), the density of the undirected graph is $4/6 = 0.66$, this means the network is *dense*. Whereas, in the second network (the directed graph), the density is $4/12 = 0.33$ that is the network is *sparse*.

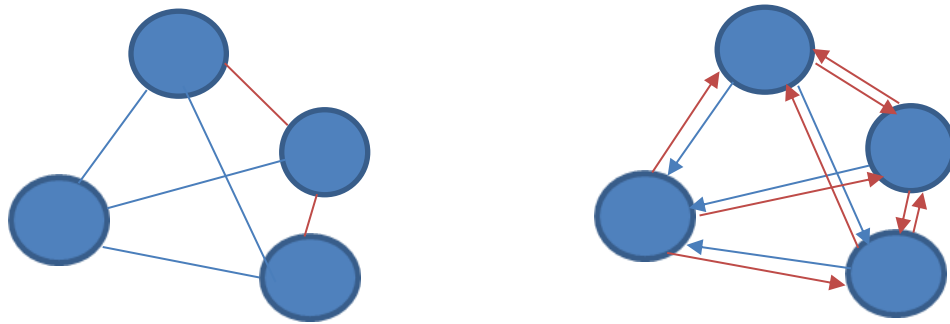


Figure 1.8: Examples of computing the density of a network.

- **Centralization:** The difference between the number of links for each node divided by maximum possible sum of differences. A centralized network will have many of its links dispersed around one or a few nodes, while a decentralized network is one in which there is little variation between the number of links each node possesses [7].
- **Closeness centrality:** “How near are you to everyone?”. The degree an individual is near all other individuals in network (directly or indirectly)[7]. A node which has high closeness requires very little travel time to get to other nodes in the network, also estimates time to hear info, influence, and point of rapid diffusion[5]. In figure(1.9), node “3” has the highest closeness in the network, it has the smallest average distance to every node in the network. On the other hand, node “1” is at the ends of the line, and so its average distance to the other nodes is at the maximum.

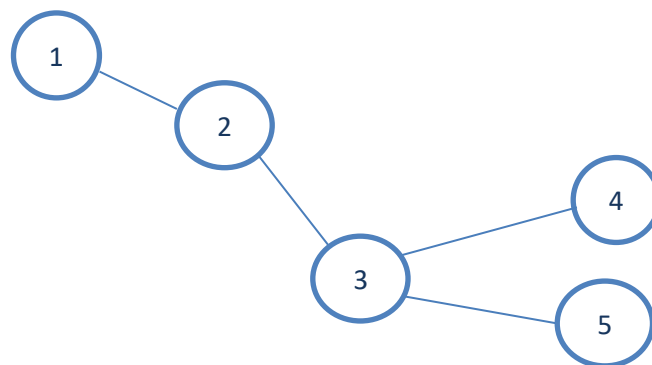


Figure 1.9: Closeness centrality of a node.

- **Degree centrality:** “How many friends do you have?”. The count of the number of ties to other actors in the network[7]. In other words, it measures the total number of edges connected to a particular vertex. In the case of Facebook, degree is the number of friends that a user has.

In the previous graph (figure (1.9)), the node “3” is the node with the highest degree in the network.

- **Betweenness centrality:** “How often do you connect other people? “. It is a measure of how many times a given node lies on the shortest path between two other nodes in a graph. Betweenness centrality is a measure of how often a given vertex lies on the shortest path between two other vertices. This can be thought of as a kind of “bridge” score[6].

The figure (1.8) shows that node “3” has the highest betweenness here, since the shortest paths from {“1”, “2”} to {“4”, “5”} need to go through the node “3”.

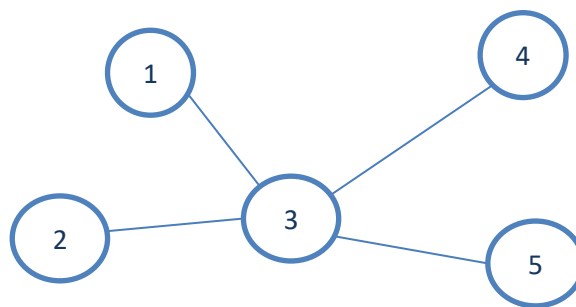


Figure 1.10: Betweenness centrality of a node.

- **Eigenvector centrality:** this is a measure of the importance of a node in a network[7]. Unlike degree centrality that gives a simple count of the number of connections a vertex has; eigenvector centrality acknowledges that not all connections are equal so

that connecting to some vertices has more benefit than connecting to others[6].

- **Clustering coefficient:** A measure of the likelihood that two associates of a node are associates themselves. A higher clustering coefficient indicates a greater 'cliquishness'[7].

8. Summary:

In recent years, the use of social network analysis has increased dramatically. Network analysis concepts and metrics are becoming widely available. The problems in collecting and analysing network data were greatly minimized by effective visualization techniques and methods. These methods allowed experts such as business managers to analyze facts so that more sophisticated and accurate decisions can be taken over a shorter timeframe.

Chapitre III. Information Diffusion Models:

1. Introduction:

The fact that the main purpose behind the creation of social media networks is to facilitate and to enhance interactions among web users, cannot be denied. Due to this, the increasing emergence of online social media networks has piqued the attention of SNA researchers in the examination of the information diffusion process and the study of social influence in social networks. The purpose of this chapter is to define influence and information propagation in social networks. Then it discusses notable diffusion models in the literature as well as the significance of analysing information flow.

2. Information diffusion in social networks:

Social networks are an effective tool for the spread of knowledge, information, thoughts, and influence among its users. And according to [19], These networks facilitate the sharing and propagation of information among members of their networks.

Information diffusion is a vast research domain and has attracted research interests from many fields, such as physics, biology, etc[20]. Also Information diffusion is a variant of an area of study in the social sciences called diffusion of innovations[21].

An innovation can be an idea, activity or object that is seen as new. And diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social system[22].

Diffusion of innovations is a theory that seeks to explain how, why, and at what rate new ideas and technology spread [23]. It explores social networks and their role in influencing the spread of new ideas and practices [7].

As mentioned by [20], the analysis of information diffusion including its capture, modelling, and visualization “help in better following events (e.g. analysing revolutionary waves), solving issues (e.g. preventing terrorist attacks, anticipating natural hazards), optimizing business performance (e.g. optimizing social marketing campaigns), etc.”

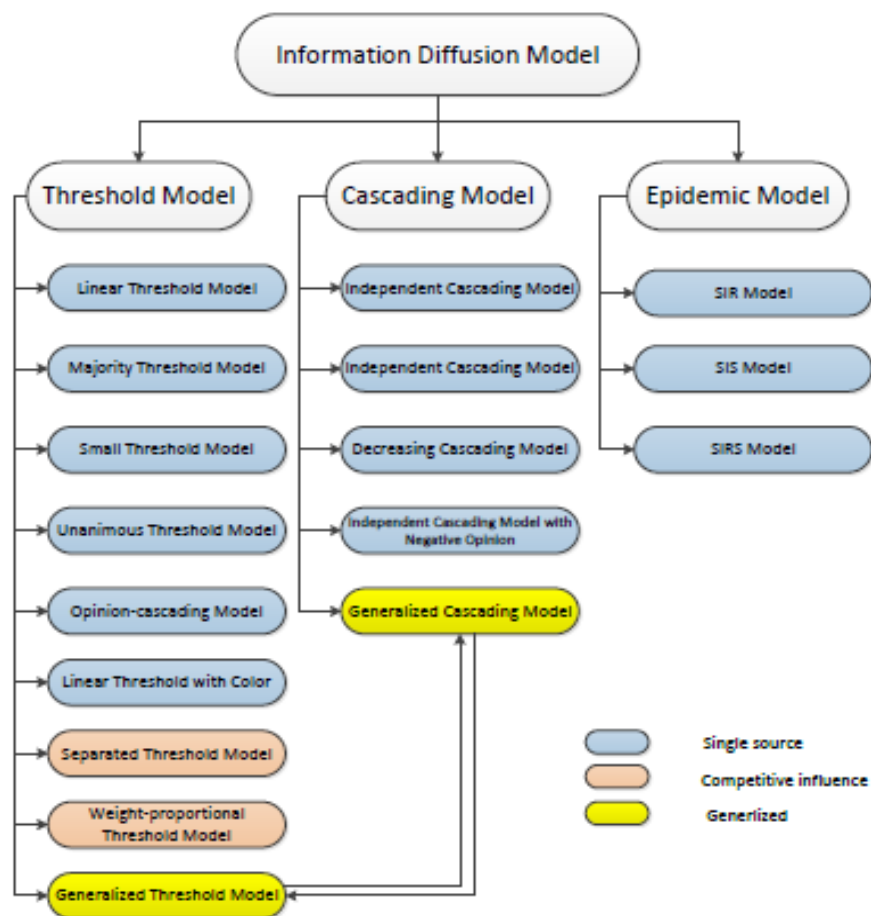


Figure 2.1: An overview of information diffusion models[17] .

Influence diffusion is the process that information propagates through certain intermediaries over time among the individuals of a social network[17].

Also, Sociologists defined social influence as a “change in an individual’s thoughts, feelings, attitudes, or behaviours that results from interaction with another individual or a group” [24].

And according to [25], “An influencer is someone who has the power to affect the purchasing decisions of others because of his or her authority, knowledge, position, or relationship with his or her audience”.

3. Diffusion models in social networks:[26] [17]

A diffusion model is the entire definition of the propagation process. It controls how spread happens. There are numerous mathematical models that have been recommended to formally model the spread of information in the network[27]. Other researchers have modelled information diffusion in social networks using physical diffusion models, such as heat diffusion processes, based on the concept that information diffusion in social networks is comparable to diffusion in physical phenomena, such as heat flowing from a high temperature area to a lower temperature area. In this model, the heat kernel models the information flow in networks (see [27] for more details about these models).

A social network is represented as a graph which can be either a directed or undirected graph according to its real application and network property[17].

The literature offers four basic approaches to modelling influence propagation in social networks: linear influence model, cascade model, threshold model and epidemic model[28][29][30]. Before we explore these models, we take a glance at the specific vocabulary in the context of information propagation in a network.

Given a network $G = (V, E)$ with V is the set of vertices, and E the set of the edges existing in the network. A vertex $v \in V$ is said to be active if the information has reached the vertex and was accepted by it. “Contaminated vertex” is the synonym of active vertex, mostly used when the study involves some viruses instead of information propagation. If the information didn’t

reach the vertex or the vertex rejected it, then the vertex is said to be inactive or non-contaminated. A vertex can change its state during the process of information diffusion from inactive to active, but not the opposite, in order to simplify the models. (In fact, in some recent researches, a vertex can become inactive again).

There must be an initial set of vertices activated to start the diffusion process targeted for initial activation. They are called '*initial adopters*' of the information. The influence of this initial set of vertices is the expected number of active vertices in the end of the diffusion process.

a) Independent Cascade Model (ICM):

These models assume that each active node v has one chance to activate each of its neighbours w with a nonzero probability $P_{v,w}$ which represents the influence of v on w . If a node w has multiple newly activated neighbours, their attempts to activate w are sequenced in a random order. Whether the activation attempt was successful or not, v will not be able to activate w again. This type of models is chronological and progressive so that it proceeds by activation steps and nodes pass only from inactive to active, but not the reverse. The process runs until no more activations are possible.

b) Linear Threshold Model (LTM)[31]:

Linear threshold model is a basic threshold model which is broadly used in analysing the information diffusion among the users. It contains the same notion as the ICM, except that the single activation attempt rule is replaced by another activation mode.

In this model, we define a social graph in which the node v is connected with its neighbour u through a weighted link whose weight is a probability $P_{uv} \geq 0$ that represents how u can affect the decisions made by v . It seems clear that the directed and weighted graph is the kind of graphs with which LTM should be applied. For all the neighbours of v , LTM suppose that:

$$\sum_{u \in N(v)} P_{uv} \leq 1$$

Where $N(v)$ represents all the neighbours of u from which it receives incoming links. Each node v is assigned a threshold θ_v which is chosen uniformly from the interval $[0, 1]$. This threshold refers to the amount of influence exerted on v by its incoming neighbours $N(v)$. In other words, θ_v is the minimum number of v 's incoming neighbours that must be active so that u will become active too.

It should be mentioned that just like the ICM, the LTM model assumes that the diffusion process happens through a number of activation steps starting with an initial set of previously activated nodes S_0 .

The diffusion process ends when, in a certain activation step s , no other inactive nodes that turn to the active state.

c) Linear Influence Model (LIM)[32]:

It differs from previous models such as LTM and ICM, which require knowledge of the social graph, and is modelled without the need for explicit network knowledge. This model ignores which node infects which other nodes, conversely choosing to focus on the rate of diffusion, which determines whether information spreads slower or faster. The influence of linear model (LIM) is defined based on the assumption that the number of newly activated nodes depends on the number of nodes activated in the past. The number of infected nodes has been formulated by their model as a function of time.

According to LIM, each node will be assigned a so-called influence function that quantifies the number of subsequent infections that can be associated to the influence of this node over time. A non-parametric formulation of LIM is given as follows:

$$V(t + 1) = \sum_{u \in A(t)} I_u(t - t_u)$$

Where

- $V(t)$ is the number of nodes that mention the information at time t .
- $I_u(l)$ represents the influence of the node u at time t after l units of time from the moment it has adopted the information.
- $A(t)$ means the set of nodes which have been activated before time t ($t_u < t$).

▪ **Modelling influence function:**

How to model $I_u(l)$?

Influence function can be modelled either by an exponential law so $I_u(l) = e^{-\lambda ul}$ or with a power law $I_u(l) = c_u l^{-\alpha u}$.

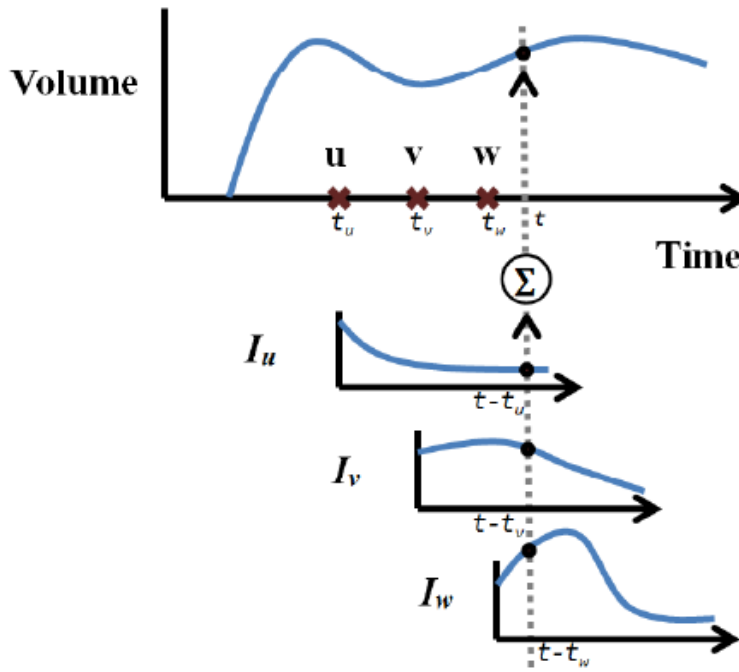


Figure 2.2: An illustration of the linear Influence Model (LIM) [32] .

Figure 2.2: represents the volume $V(t)$ over time, and t_u , t_v , and t_w denote the times when

nodes, u , v and w , got infected. After the nodes got infected, they each influence additional $I_u(t - t_u)$, $I_v(t - t_v)$ and $I_w(t - t_w)$ infections at time t . So, the volume $V(t)$ at time t is the sum of the influences of the three nodes.

d) Epidemic Model:

An epidemic model describes the transmission of contagious disease through individuals. In the recent century, it has been widely used to model computer virus infections and information propagations such as news and rumours[17].

- **SIR Model**[33]: (Susceptible-Infectious-Recovered) model first proposed by Kermack and McKendrick (1927) [34]. This model is the one most commonly used class of the continuous-time epidemic models. A node in a network goes through three potential stages during the course of the epidemic: (1) *Susceptible* to infection from its neighbours, (2) *Infectious* if it has caught the disease and has some probability of infecting each of its susceptible neighbours, (3) *Recovered* or *Removed* from consideration (in case of death) after the full infectious period (see figure 2.3). Given a directed graph representing the contact network; so, an edge pointing from node v to node w in the graph means that if v becomes infected at some point, the disease has the potential to spread directly to w . each node has the potential to go through the Susceptible-Infectious- Removed cycle (S-I-R). The progress of the epidemic is controlled by the contact network structure and by two additional quantities: p (the probability of contagion) and t_i (the length of the infection). Figure (2.3) shows an example of the SIR model in a small network through successive steps. In each step, shaded nodes with dark

borders are in the I state and shaded nodes with thin borders are in the R state.

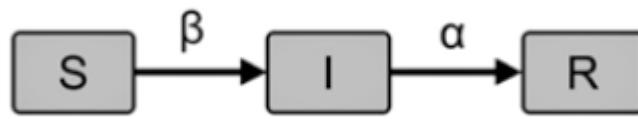


Figure 2.3: SIR model showing the three phases.

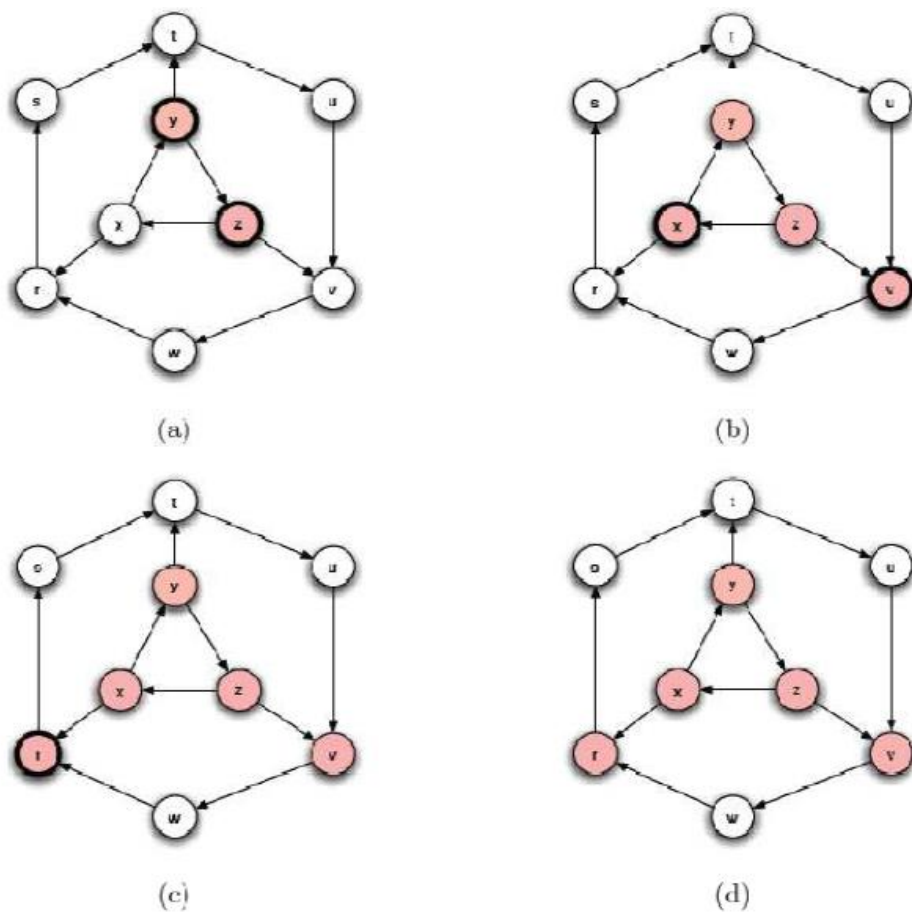


Figure 2.4: Illustration of the SIR Epidemic model[26].

4. Summary:

There are an infinite number of cases in which people are impacted by others. In this chapter, we concentrated on providing a concise and obvious overview of the most common influence and information diffusion models which are: the epidemic model, independent cascade model; linear threshold model and linear influence model. These models aid analysts in comprehending the flow of information and, in particular, exerting control over the social network's connection structure.

In the following chapter, I will clarify with more details the models to be used in my experiments: the independent cascade model, and the linear threshold model.

Chapitre IV.Implementation:

1. Introduction:

In the earlier chapter, we discussed the general concepts of influence and modelling information diffusion.

In the following; We will explain the diffusion process in wider detail, then we focus on modelling this process using Independent Cascade Model (ICM) and Linear Threshold Model (LTM) that we use in the experiments. Later, we describe the tools we used and the choices we made to achieve the main goals.

2. Information cascades:

On social networks, users regularly repost content (texts, photographs, videos, etc.) already uploaded by other users. This content will be immediately received by their neighbours (i.e., contacts like friends). We say that an information cascade occurs every time that content moves from a circle of people to another one. The items involved in the cascade process should be connected by a network, and they see their immediate neighbours.

In general, a cascade of information might develop when people sequentially ignore their own knowledge (which comes from their ambiguous opinions) and draw conclusions from the acts of their predecessors [35].

Three observations are found in the information cascade that are [36]:

First observation is that "*cascades can be wrong*", That means do they do (*Following the Crowd*) for example, you went to the store to buy something

specific then you found a big crowd around other things but without thinking, you bought it and when you came home, you didn't like it, in this case, following the crowd was a wrong choice for you.

One must also notice that "*cascades can be based on very little information*", because people ignore their own information in the beginning of a cascade, which means the behaviour of the population is influenced only by the pre-cascade information. This means that if a cascade starts relatively quickly in a large population, most of the private information that is collectively available to the population is not being used.

Finally, "*cascades are fragile*". since cascades can be based on little information, they can stop as easy as they can start. One manifestation of this that people who receive slightly superior information can overturn even long-lived cascades.

3. Modelling information diffusion:

In social networks and especially in the social media, researchers examined information cascades. They proposed many mathematical models to represent the dissemination of information in the network explicitly. In earlier parts, I mentioned four renowned models: the Epidemic Model, Linear Influence Model (LIM), Independent Cascade Model (ICM) and Linear Threshold Model (LTM). To achieve the goals of this work, I chose two basic diffusion models: the Linear Threshold Model and Independent Cascade Model.

In implementing the chosen models, I focus on the following settings: given a network presented by a directed graph $G = (V, E)$, each vertex $v \in V$ can be either active or inactive. Each inactive vertex tends to become active, and it can switch from being inactive to being active, but do not switch in the other direction. Consequently, the cascading process will appear as follows: given an initial set of active vertices; while time spreads out, more of an inactive vertex v 's neighbours become active which may at some moment make that vertex active. In turn v can trigger the same

choice or action by other vertices to which it is related to adopt the same decision or action. The influence of this initial set is the expectation number of active vertices by the end of the diffusion process running.

- **LTM:** Many ways for information dissemination have been proposed as I have previously mentioned. One basic approach is based on the use of node-specific thresholds. The first to suggest threshold models in sociology was Granovetter[37]: *"the key concept is that of threshold: the number proportion of others who must make one decision before a given actor does so"*.

The adopted threshold model I first implement is: the Linear Threshold Model. In this model, a node V is influenced by each neighbour W according to a weight b_{vw} . Each vertex V chooses a threshold uniformly at random from the interval $[0,1]$. After choosing the thresholds randomly, an initial set of active nodes A_0 is also chosen.

As explained before, the diffusion process unfolds in discrete steps: in step s : all vertices that were active in step $(s-1)$ remain active, and new vertex v is activated if the total weight of its active neighbours is at least T_v . The information stops propagating when there are no more inactive vertices that can become active.

$$\sum_{w \in N_v} b_{vw} \geq T_v$$

The random choice of thresholds T every time the diffusion process is carried out, fill the lack of information about the network and the relations between its actors. These thresholds are established for all vertices of the network in certain approaches to a particular value (typically 0.5).

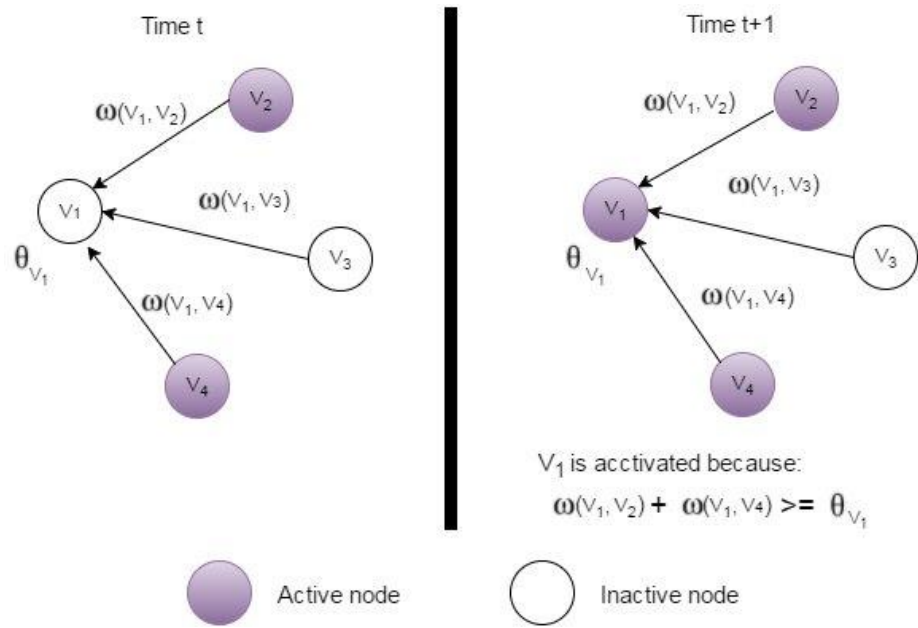


Figure 3.1: Linear Threshold Model example[38].

This is how LTM works, and here is its algorithm:

The LTM algorithm

1. $S_0 \neq \{\}$; initial set of nodes selected for launching the diffusion.
2. $i = 0$;
3. *thresholding* (V); Attributing to each node v a threshold θ_v from the interval $[0,1]$.
4. **While** $i = 0$ ou $(S_{i-1} \neq A_i, i \geq 1)$ **Do**
5. $A_{i+1} = A_i$;
6. $inactif = V - A_i$; The set of inactif nodes
7. **For** $v \in inactif$ **Do**
8. **If** $\sum_{u \text{ an incoming neighbor of } v, u \in A_i} P_{uv} \geq \theta_v$ **Then**
9. Activate (v);
10. $A_{i+1} = A_{i+1} \cup \{v\}$;
11. **End if**
12. **End For**
13. $i = i + 1$;
14. **End While**

- **ICM:** I adopt the simplest and most studied diffusion model: the Independent Cascade Model. As with LTM, this model starts with an initial set of active vertices A_0 , and the process unfolds in discrete steps according to the following randomized rule. When vertex v first becomes active in step s , it is given a single chance

to activate each of its inactive neighbours w with a probability p_{vw} to succeed. If v succeeds to activate w in step $s+1$, then w will try to activate its inactive neighbours too. Otherwise, v cannot make any further attempts to activate w in subsequent rounds. In case that w has many new active neighbours, they attempt to active w in an arbitrary order.

The diffusion process ends when there is no more inactive vertices that can be switch to become active at a step s .

The algorithm of this model can be written as follows:

The ICM algorithm

```

1.  $S_0 \neq \{\}$ ;
2.  $i = 0$ ;
3. While  $S_i \neq \{\}$  Do
4.    $i = i + 1$ ;
5.    $A_i = \{\}$ ;
6.   For  $v \in A_{i-1}$  Do
7.     For  $w$  a neighbor of  $v$ ,  $w \notin \bigcup_{j=0}^i A_j$  Do
8.        $r = \text{random value} \in [0,1]$ 
9.       If  $r < P_{v,w}$  Then
10.        Activate ( $w$ );
11.         $A_i = A_i \cup \{w\}$ ;
12.       End If
13.     End For
14.   End For
15. End While

```

4. How to choose initial active set:

Another important step in the implementation of the chosen models is “choosing the initial active set”. Many researchers focused on the problem of

maximizing the influence or the spread of information in social networks by choosing the best initial adopters and the most influential ones[39].

The problem is posed as follows: "if you want to trigger a large cascade of adoptions of a new product or innovation, you need first to convince a subset of individuals to do so. In this case, which set of individuals (nodes) should you target initially?"

initial active set is chosen based on 3 criteria: The Degree metric, betweenness centrality, and randomly. Next, every choice is described with details.

× **Degree centrality:**

The degree centrality is a count of total number of connections linked to a node. In real world networks, the degree centrality measures how many people a person can reach directly in the network. The higher the degree is, the more important the node is. there are two measures of degree. In-degree represents how many links those point inward at a node, while the out-degree is the number of links those point outward to other nodes.

The dilemma now is: how to choose the nodes? Based on the high in-degree or out-degree? Their sum? Or even the difference between them? To better understand the situation here, an example is needed to be illustrated.

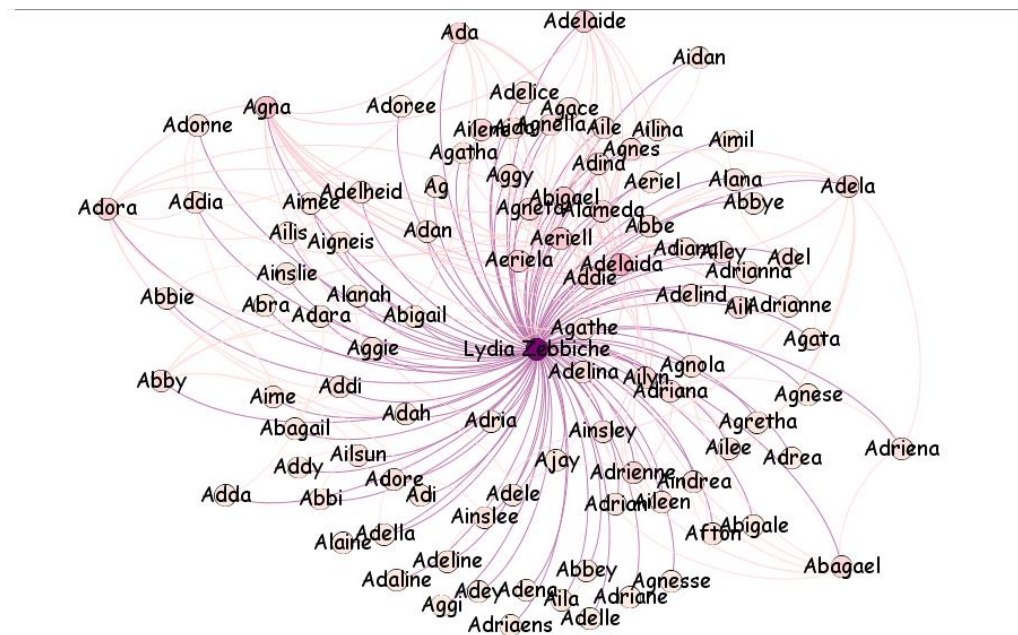


Figure 3.2: Degree appearance on my graph

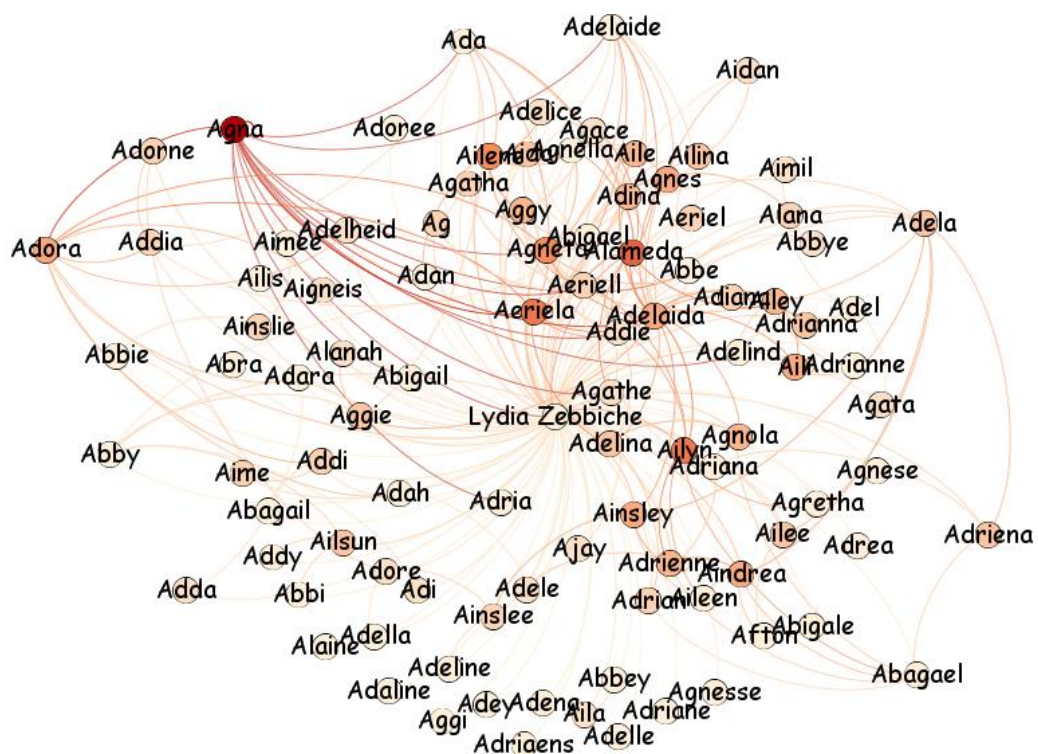


Figure 3.3: In-degree appearance on my graph

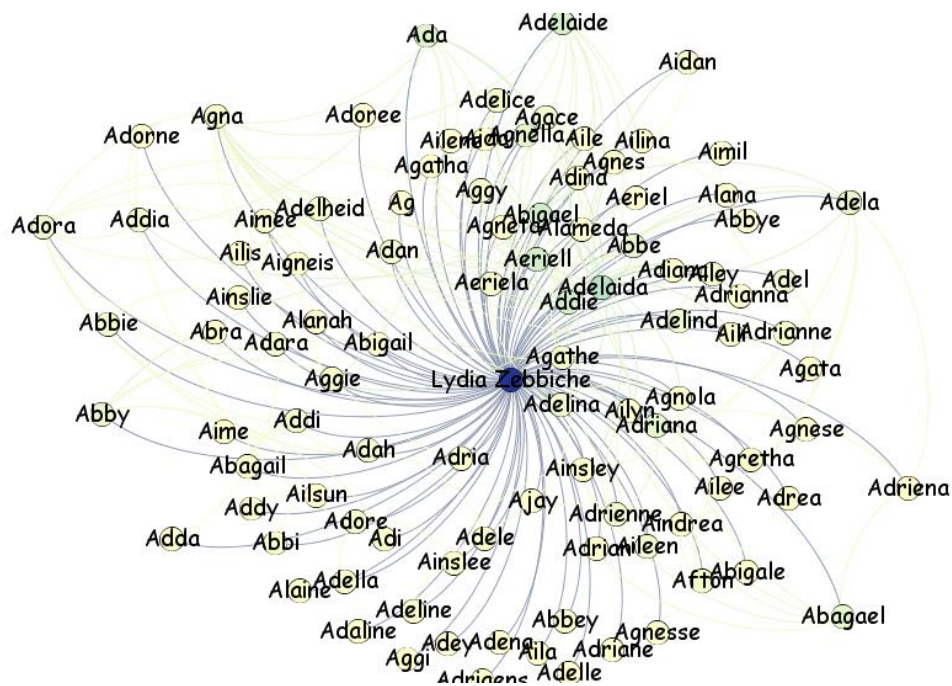


Figure 3.4: Out-degree appearance on my graph

Id	Label	Interval	In-Deg...	Out-Degree	Degree
0	Lydia Zebbiche		1	100	101
25	Adelaida		6	18	24
66	Agna		15	1	16
56	Aeriell		3	11	14
21	Addie		2	11	13
9	Abigael		1	12	13
55	Aeriela		9	2	11
26	Adelaide		1	10	11
40	Adora		6	4	10
97	Alameda		10	0	10
24	Adela		4	5	9
53	Adrienne		5	4	9
13	Ada		1	8	9
87	Ailyn		9	0	9
71	Agneta		8	0	8
81	Ailene		8	0	8
68	Agnes		6	1	7
30	Adelina		4	3	7
82	Ailey		5	2	7
1	Abagael		1	6	7
3	Abbe		1	6	7
48	Adriana		1	6	7
67	Agnella		1	6	7
83	Aili		6	1	7

Figure 3.5: list of high degree that I used it in my experiment

× **Betweenness centrality:**

betweenness centrality metric assesses the importance of a node based on its position in the network. It is the number of shortest paths from all nodes to all others that pass through that node. The idea is that actors who are “between” other actors, and on whom other actors must depend to reach others are more important in a social network.

Betweenness centrality is determined in two main steps[40]: first the length and the number of shortest paths between all pairs of nodes must be computed. The second step is summing all pair-dependencies. The pair-dependency of a node v is the ratio of shortest paths between s and t that v lies on.

A path is defined from $s \in V$ to $t \in V$ as a sequence of vertices and edges, beginning with s and ending with t . the length of a path is the sum of the weight of its edges. Let σ_{st} denotes the number of shortest paths from s to t , where $\sigma_{ss} = 1$ by convention. And let $\sigma_{st}(v)$ denotes the number of shortest paths from s to t that some $v \in V$ lies on. The betweenness centrality is defined as follows:

$$C_B = \sum_{s \neq v \neq t \in V} \delta_{st}(v)$$

Where $\delta_{st}(v)$ is the pair-dependency of vertex v , it is computed as follows:

$$\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Id	Label	Interval	In-Deg...	Out-D...	Degree	Betweenness Centrality
25	Adelaida		6	18	24	66.0
56	Aeriell		3	11	14	48.0
24	Adela		4	5	9	43.0
21	Addie		2	11	13	30.5
40	Adora		6	4	10	30.0
68	Agnes		6	1	7	10.0
53	Adrienne		5	4	9	9.0
30	Adelina		4	3	7	8.0
19	Addi		3	2	5	7.0
62	Agathe		2	3	5	7.0
20	Addia		3	2	5	6.0
64	Aggie		4	1	5	6.0
28	Adelheid		2	1	3	5.0
27	Adele		2	1	3	4.0
39	Adina		5	1	6	4.0
75	Aidan		2	1	3	4.0
82	Ailey		5	2	7	4.0
65	Aggy		5	1	6	3.5
41	Adore		2	2	4	3.0
92	Ainslee		3	1	4	3.0
8	Abbye		2	1	3	1.0
29	Adelice		2	2	4	1.0
55	Aeriela		9	2	11	1.0
0	Lydia Zebbiche		1	100	101	0.0

Figure 3.6: list of high betweenness that I used it in my experiment.

× Random:

Lastly, the same algorithms are tested as a fundamental and understandable choice if the initial active set is randomly selected without a particular requirement. Lastly, the same algorithms are tested as a fundamental and understandable choice if the initial active set is randomly selected without a particular requirement.

initial activation nodes
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95

Figure 3.7: list of randomly choose.

When it comes to the number of initial active nodes, there is no general rule to make this choice. From the literature, the size of the initial active set is usually between 30 to 50 nodes. This depends on the size of network, its type, and the domain of the experiments. In this work the size of initial active set is chosen between 1 to 20 nodes.

5. Experiments tools:

× Gephi [41]:

In my experiments, I've used "Gephi0.9.2" tool. Gephi is an open-source software for network visualization and analysis. It uses a 3D render engine to display large graphs in real-time and to speed up the exploration. Gephi combines built-in functionalities and

flexible architecture to: Explore, analyse spatialize, filter, cluster, manipulate and export all types of networks.

It is one of the most useful programs to explore and understand graphs that consist of nodes and links between them, or so-called graph, where the user can manipulation of structures, shapes and colours, to detect the properties of the network, and explore their hidden properties.

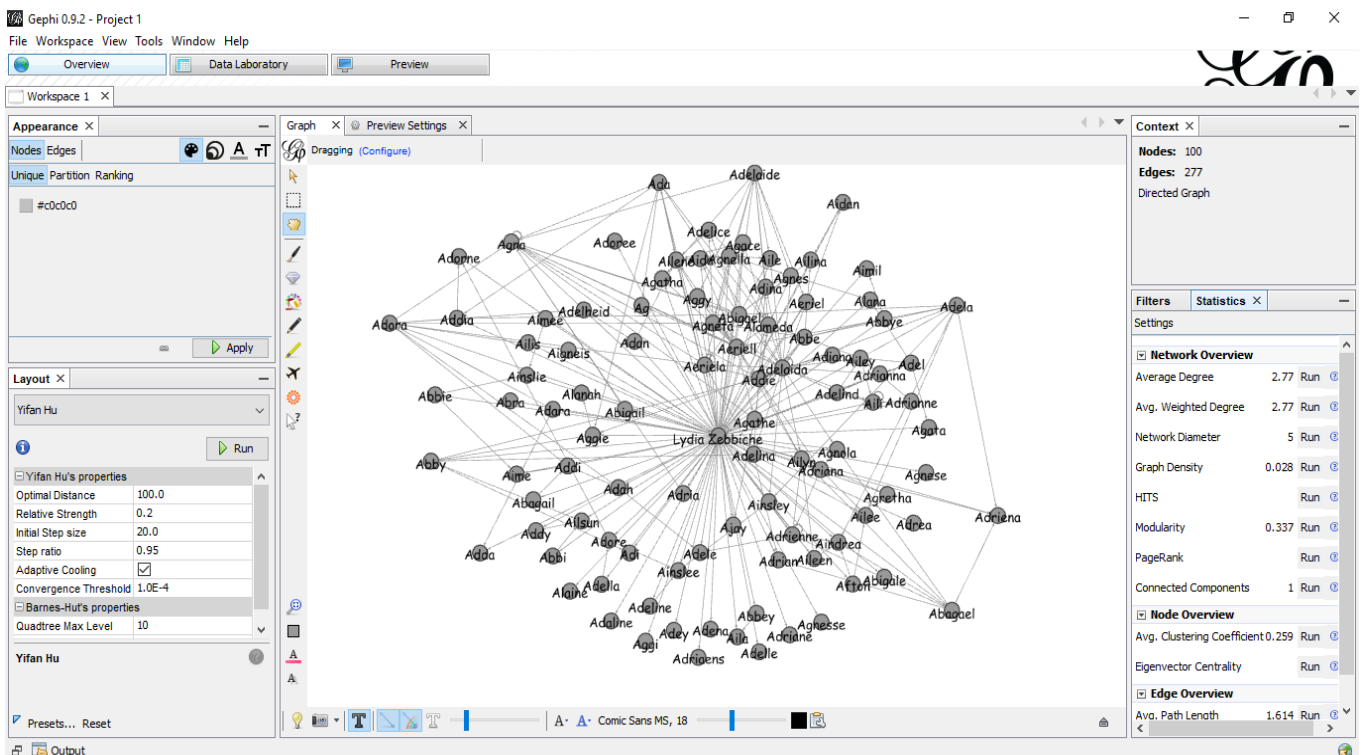


Figure 3.8: The complete Gephi workspace window

× **Python**[42]:

Gephi doesn't contain any information diffusion model. For that reason, I've used Python for programming the discussed models. Python is a high-level programming language designed to be easy to read and simple to implement. It is open source, which means it is free to use, even for commercial applications.

I used Ndlib library to call both models ICM and LTM , figure 4.8 shows LTM model and figure 4.9 shows ICM models.

```
out_LTM = []
for ii in range(0,115,5):
    initialSetNodes= ((100*(ii))/len(list(g.nodes())) )/100
    print(initialSetNodes)
    model = ep.ThresholdModel(g)

    # Model Configuration
    config = mc.Configuration()
    config.add_model_parameter('fraction_infected', initialSetNodes)

    # Setting node parameters
    threshold = 0.5
    for i in g.nodes():
        config.add_node_configuration("threshold", i, threshold)

    model.set_initial_status(config)

    # Simulation execution
    iterations = model.iteration_bunch(999)
    #print(iterations)
    colors = []
    ColorLegend = {'Susceptible': 0, 'Infected': 1, 'removed': 2}
    some_colors = ['orange', 'green', 'yellow', 'purple', 'magenta', 'pink']
    k = 0
    activ = 0
    for iterr in iterations:

        colors = ["blue" for x in g.nodes()]

        for stat in iterr["status"]:

            val = iterr["status"]["{}".format(stat)]

            if val == 1:
                activ+=1

            if val > 0:
                index = list(g.nodes()).index(stat)
                colors[index] = some_colors[val-1]

        nx.draw(g, with_labels = True, font_color="red", node_color=colors)

        plt.draw()

        plt.savefig("figLTM/iteration_{}.jpg".format(k), dpi=600)
        plt.clf()

        k+=1
        #print(iterr, activ)
    out_LTM.append({"initial activation nodes ":ii, "active_nodes":activ})
    #print(out_LTM, ii)

pd.DataFrame(out_LTM).to_excel("LTM.xlsx", engine="openpyxl")
```

Figure 3.9: LTM by Python.

```

out_Indepednet = []
for ii in range(0,115,5):
    initialSetNodes= ((100*(ii))/len(list(g.nodes())) )/100
    model = ep.IndependentCascadesModel(g)

    # Model Configuration
    config = mc.Configuration()

    config.add_model_parameter('fraction_infected',initialSetNodes )

    # Setting the edge parameters
    threshold = 0.5
    for e in g.edges():
        config.add_edge_configuration("threshold", e, threshold)

    model.set_initial_status(config)

    # Simulation execution
    iterations = model.iteration_bunch(999)
    #print(iterations)

    trends = model.build_trends(iterations)

    # Visualization
    viz = DiffusionTrend(model, trends)
    viz.plot("diffusion.pdf")

    colors = []
    ColorLegend = {'Susceptible': 0,'Infected': 1,'removed': 2}
    some_colors = ['orange','green','yellow','purple','magenta','pink']
    k = 0
    activ = 0
    for iterr in iterations:

        colors = ["blue" for x in g.nodes()]
        for stat in iterr["status"]:

            val = iterr["status"][{"{0}".format(stat)}]
            if val == 1:
                activ+=1

            if val > 0:
                index = list(g.nodes()).index(stat)
                colors[index] = some_colors[val-1]

        nx.draw(g,with_labels = True,font_color="red",node_color=colors)

        plt.draw()

        plt.savefig("fig/iteration_{0}.jpg".format(k),dpi=600)
        plt.clf()

        k+=1
    out_Indepednet.append({"initial activation nodes ":ii ,"active_nodes":activ})

pd.DataFrame(out_Indepednet).to_excel("IndependentCascadesModel.xlsx",engine="openpyxl")

```

Figure 3.10: ICM by Python.

6. Datasets:

⌘ Facebook (user's list):

Facebook is an online social media and social networking service that allows users to post images, photos, videos, files and documents, exchange messages, join and create groups and use different of applications. It gives people the power to build community and bring the world closer together. People use Facebook to stay connected with friends and family, to discover what's going on in the world.

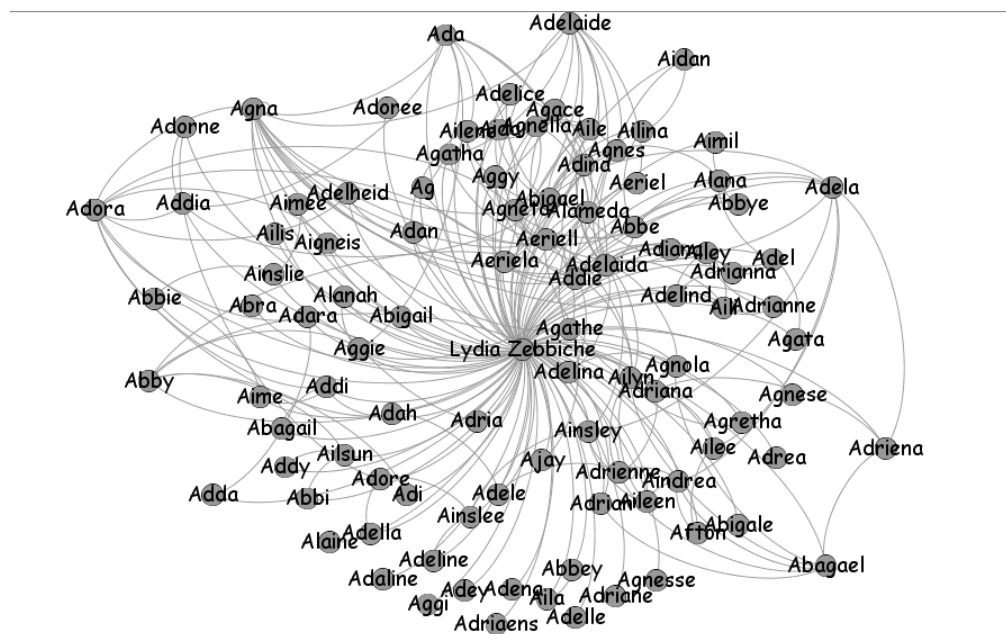


Figure 3.11: Graph of my datasets

I experimented with a variety of Gephi layouts before settling on the layout called Yifan Hu which formed the network depicted in Figure 4.10.

Figure 4.11 represents two subgraphs which illustrated the degree between node "Abagael" to "Lydia Zebbiche" (graph A) and "Adelaid" to "Lydia Zebbiche" (graph B).

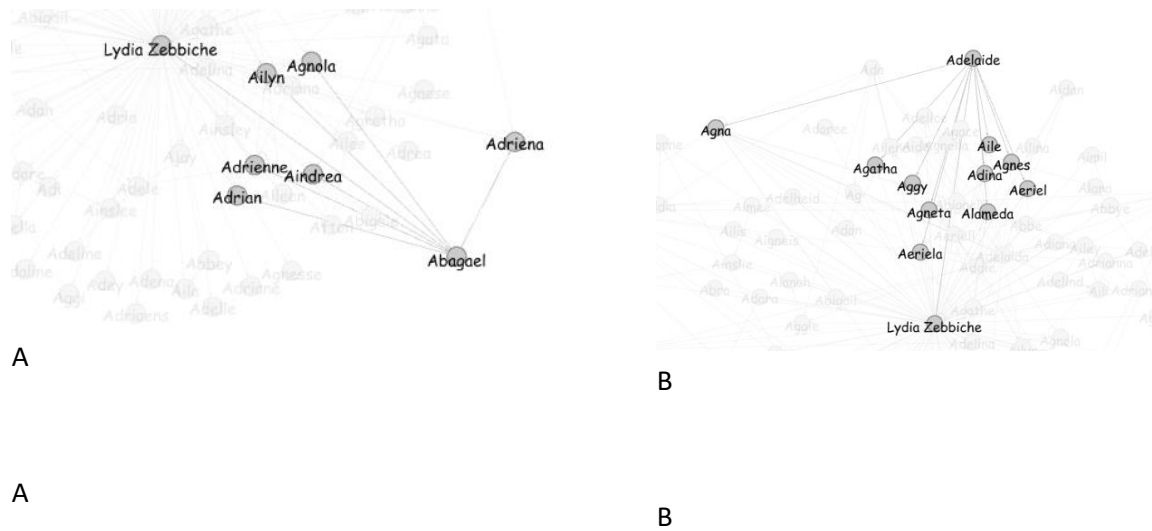


Figure 3.12: Subgraphs of different degrees.

The network has 277 edges and 100 nodes represented by the edges list figure 4.12

Data Table		
Nodes Edges Configuration Add node Add edge Search/Replace Import Spreadsheet Export table More actions		
Source	Target	Type
33	41	Directed
39	68	Directed
40	55	Directed
40	66	Directed
40	71	Directed
40	76	Directed
41	43	Directed
41	92	Directed
45	66	Directed
48	52	Directed
48	53	Directed
48	56	Directed
48	72	Directed
48	79	Directed
48	87	Directed
51	82	Directed
51	83	Directed
53	53	Directed
53	87	Directed
53	91	Directed
53	93	Directed
55	55	Directed
55	66	Directed
56	58	Directed

Figure 3.13: The edges list shows relationships among the nodes.

7. Summary:

This chapter was divided into three sections. In the first section, we showed how information may spread in social networks by forming information cascades. The second section dealt with the implementation of the models we picked to assess information transmission on social media. Whereas, the section clarified the tool, programming language, and datasets we utilized.

In the following chapter, the details of the experiments, as well as their outcomes and conclusions, are presented.

Chapitre V. Experiment's Results:

1. Introduction:

In this chapter, we examine results from the experiments for estimating the number of infected nodes in a social media and to answer the questions asked in the first chapter. we begin by understanding each network structure and its metrics in the first part. I use plots and metrics visualization to make the analysis more understandable. In the second part, I use the chosen models to demonstrate the various diffusion process testing phases in the networks. Their results are given and construed one by one, with some examples. Lastly, the final part analyses the results and makes generalized conclusions.

2. Network's structure and metrics analysis:

The social networks used in my experiments, namely Facebook user's that has size, type, and its structure. In table 5.1 I resume the network structure characteristics.

	Facebook user's friends
Graph type	Directed
Egocentric network	Yes
Type of relationship	Implicit
Number of Nodes	100
Number of Edges	277
Graph density	0.028
Connected component	1
Maximum number of nodes in connected component	100
Maximum number of edges in connected component	277
Diameter	5
Average Path length	1.614141414141414

Table 4.1: The structure characteristics of Facebook users.

The centrality metrics are summarized in table (4.2) and their distribution plots are drawn on figure (5.2) for Facebook user's network. In the network (Facebook user's), one can notice that most nodes have a null value for out-degree, and their in-degree equals to 1. Only few nodes, have significant values, where the highest in-degree value is only 9. Therefore, the Max in-degree and out-degree is 9 and 83 respectively.

Average Degree	2.770	
Out-degree	Min	Max
	0	100
In-degree	1	15
betweenness	00	66

Table 4.2: The centrality metrics.

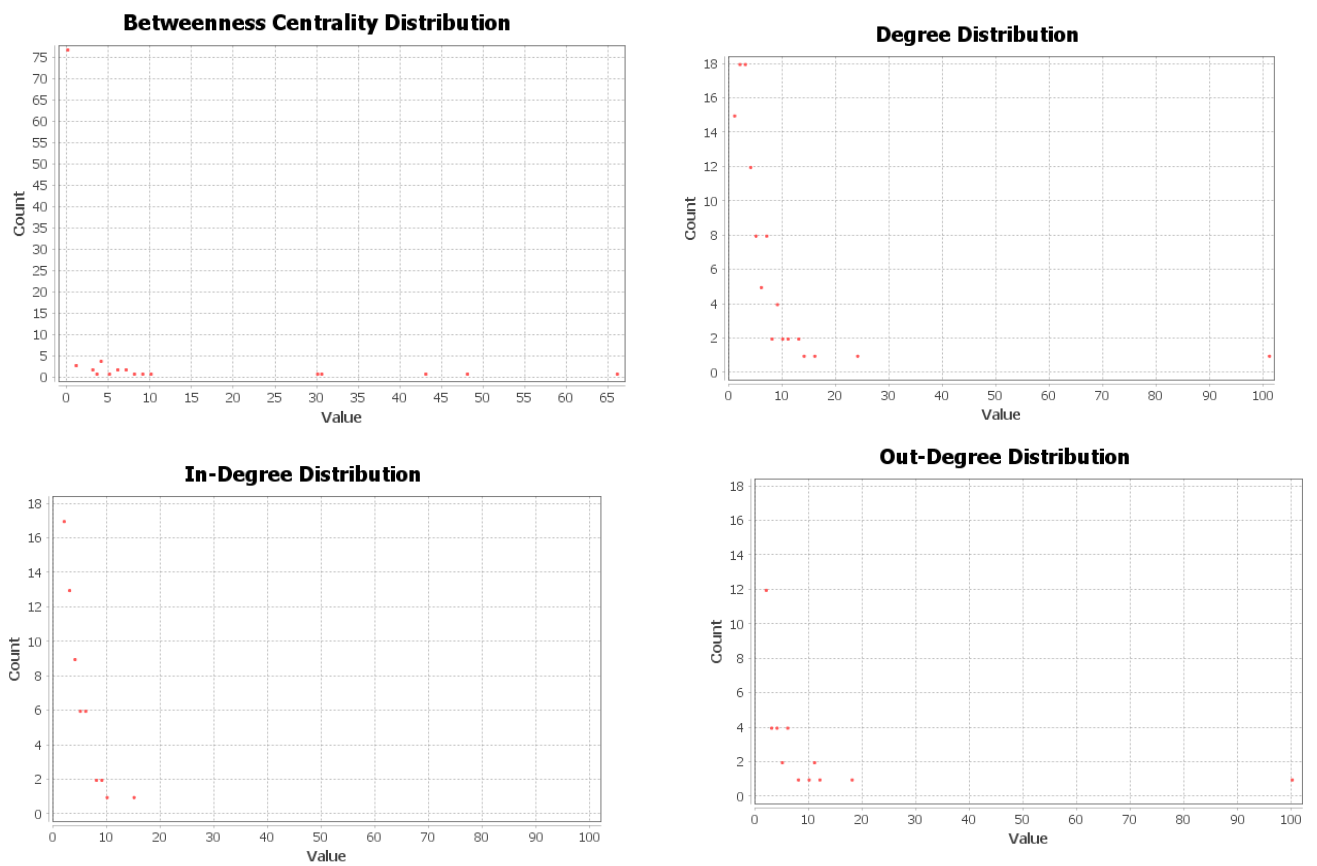


Figure 4.1: Facebook user's centrality metrics distribution plots.

Number of Initial activation nodes	Randomly		Degree		Betweenness	
	LTM	ICM	LTM	ICM	LTM	ICM
1	0	1	0.357142857	0.428571429	0.125	0.142857143
5	95	24	0.068455134	0.120879121	1.361702128	0.214285714
10	90	12	0.061016949	0.074596774	2.966666667	0.75
15	85	22	0.05967366	0.079281184	4.136363636	2.454545455
20	80	19	0.05541347	0.071977638	4.014492754	3.759259259

Table 4.3: The nodes activated before 1000 runs (average).

The results are presented in this section from dataset. For each algorithm ICM & LTM, the propagation process was run 1000 times for every initial set $A_0 \in \{1, 5, 10, 15, \text{ and } 20\}$, then the average of the size of active nodes after each run was computed. This average is considered to be the influence of the initial set.

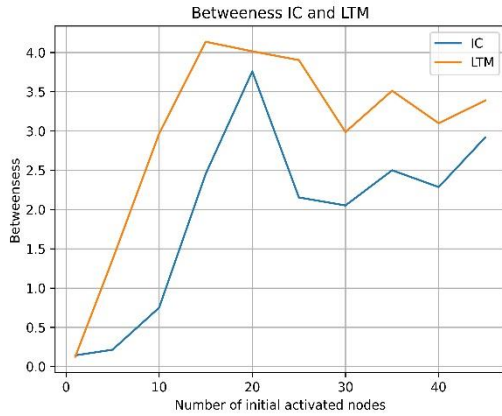


Figure 4.2: The Results of betweenness

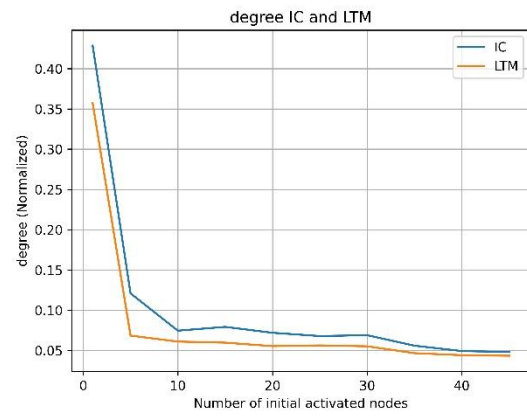


Figure 4.3: The Results of Degree

Each algorithm allows us to analyse the diffusion process in order to gain new insights and conclusions. We conducted three different analyses, each with its own set of criteria for determining results. In order to review our analysis, we used both the ICM and LTM algorithms.

In the first part, we analysed the number of active nodes using a random criterion. However, increased number of active nodes did not reflect the results because the criteria on which the test was based were selected randomly.

On the second stage, like the first point. But the criteria we had been specified were betweenness. We have gotten positive results with minor improvements utilizing ICM and LTM algorithms, so the influencing node determination will certainly be affected by this criterion.

In the third, we utilized degree as criteria instead of betweenness, but in contrast to the second phase there was also a linear result, we got less nodes at the conclusion of the research.

3. Summary:

This chapter introduces the outcomes of my work's experiments in three main areas section. The first part uses Gephi to analyse the structure and indicators of the two used data sets. In the second part we discussed and examined the various information diffusion prediction findings for each network. we also visualized other diffusion process examples to help me understand the results and how they were analysed. Finally, we talked about the results of the experiments.

A general conclusion will follow, which will summarize all of my work and hint to future study and perspectives.

General conclusion:

Human social life has been closely connected to the Internet and social media. And due to the extensive Internet disponibility and the evolution of information and communication technologies (ICT) tools that simplified, quicker and even more reliable user interaction. Social media, which has recently spread to social networking sites, is considered the latest development on the Internet. Social media facilitates the creation and exchange of information, ideas, professional skills, and other forms of expression through virtual networks.

Analysing the information diffusion on social media and the influence of users in a network is one of the most important research recently. Thus, this thesis utilizes information dissemination models to forecast the number of infected nodes and the capacity of people to distribute information across networks in social media networks. The utilized network was a Facebook page that involves node-to-node interactions.

The prediction of information propagation on this network was done by implementing two most outstanding models from the literature which are: the Linear Threshold Model (LTM) and the Independent Cascade Model (ICM). the analysis of this dissemination was based on the structure (type, size, diameter, density, connected components). Both algorithms of ICM & LTM are applied criteria which are: nodes with high overall degree, betweenness centrality and random were chosen.

However, certain experimental results were extremely unexpected and surprising when we saw from the results on the same Network that the LTM model offers more active nodes than ICM owing to the fact that the former depends on a node that is impacted by every neighbour and random threshold (each edge has its own value). Eventually, our research findings demonstrate that the selection of initial set of nodes affects significantly the diffusion process.

Bibliography:

- [1] S. A. Shrikant Dadhich, "What is the function of social media platforms? - Quora." <https://www.quora.com/What-is-the-function-of-social-media-platforms> (accessed Mar. 21, 2021).
- [2] L. S. Cohen, "Is There A Difference Between Social Media And Social Networking? Retrieved December 4, 2010." 2009.
- [3] S. Edosomwan, "SocialMedia-JAME," vol. 16, no. 3, 2011, [Online]. Available: www.LunarStorm.se.
- [4] S. K. Sajithra K, "Social Media – History and Components," *IOSR J. Bus. Manag.*, vol. 7, no. 1, pp. 69–74, 2013, doi: 10.9790/487x-0716974.
- [5] C. Science and C. Science, "Influence Analysis In Social Nets."
- [6] D. Hansen, B. Shneiderman, and M. A. Smith, *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann, 2010.
- [7] A. Kosorukoff and D. L. Passmore, *Social Network Analysis: Theory and Applications*. Passmore, D. L, 2011.
- [8] "Social Network Analysis: An Introduction by Orgnet,LLC." <http://www.orgnet.com/sna.html> (accessed May 07, 2021).
- [9] F. Bonchi, C. Castillo, A. Gionis, and A. Jaimes, "Social network analysis and mining for business applications," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–37, 2011, doi: 10.1145/1961189.1961194.
- [10] P. Anklam, *Net Work. A practical guide to creating and sustaining networks at Work and in the World*. 2007.
- [11] D. Camacho, Á. Panizo-Lledot, G. Bello-Orgaz, A. Gonzalez-Pardo, and E. Cambria, "The four dimensions of social network analysis: An overview of research methods, applications, and software tools," *arXiv*, no. February, 2020.
- [12] N. Green, "Functional polycentricity: A formal definition in terms of social network analysis," *Urban Stud.*, vol. 44, no. 11, pp. 2077–2103, 2007, doi: 10.1080/00420980701518941.

- [13] B. D. Easley, "2 Graphs," vol. 1, pp. 5–42, 2013, doi: 10.1090/text/024/02.
- [14] L. C. Freeman, "Visualizing social networks," *J. Soc. Struct.*, vol. 1, no. 1, p. 4, 2000.
- [15] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, "A framework for community identification in dynamic social networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 717–726.
- [16] "(PDF) Social Media and Academic Performance: Does Facebook Activity Relate to Good Grades?"
https://www.researchgate.net/publication/273450787_Social_Media_and_Academic_Performance_Does_Facebook_Activity_Relate_to_Good_Grades/figures?lo=1 (accessed May 17, 2021).
- [17] H. Zhang, S. Mishra, and M. T. Thai, "Recent advances in information diffusion and influence maximization of complex social networks," *Opportunistic Mob. Soc. Networks*, pp. 37–70, 2014, doi: 10.1201/b17231.
- [18] "Introduction to Social Network Methods: Chapter 8: More Properties of Networks and Actors." https://faculty.ucr.edu/~hanneman/nettext/C8_Embedding.html (accessed May 15, 2021).
- [19] B. Furht, *Handbook of social network technologies and applications*. Springer Science & Business Media, 2010.
- [20] A. Guille and H. Hacid, "2013 Diffusion Survey."
- [21] "Precursors of Panmodernism, Parts 2 thru 4." <http://panmodern.com/precursors.html> (accessed May 22, 2021).
- [22] E. M. Rogers, A. Singhal, and M. M. Quinlan, *Diffusion of innovations*. 2019.
- [23] "Diffusion of innovations - Wikipedia."
https://en.wikipedia.org/w/index.php?title=Diffusion_of_innovations (accessed May 22, 2021).
- [24] "Affect, Behavior, and Cognition – Principles of Social Psychology – 1st International Edition."
<https://opentextbc.ca/socialpsychology/chapter/affect-behavior-and-cognition/> (accessed Jun. 07, 2021).
- [25] "What is an Influencer? - Social Media Influencers Defined [Updated 2021]."

- <https://influencemarketinghub.com/what-is-an-influencer/> (accessed Jun. 07, 2021).
- [26] H. Loucif, A. Boubetra, and S. Akrouf, "A Simplistic Model for Identifying Prominent Web Users in Directed Multiplex Social Networks: A Case Study using Twitter Networks," *New Rev. Hypermedia Multimed. J.*, vol. 22, pp. 287–302, 2016.
 - [27] S. M. Ross, "Introduction to Probability Models, 11th edn." Elsevier/Academic Press, Amsterdam, 2014.
 - [28] J. Vosecky, D. Jiang, K. W. T. Leung, and W. Ng, "Dynamic multi-faceted topic discovery in twitter," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 879–884, 2013, doi: 10.1145/2505515.2505593.
 - [29] F. Celli, F. M. L. Di Lascio, M. Magnani, B. Pacelli, and L. Rossi, "Social network data and practices: The case of friendfeed," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6007 LNCS, no. May 2014, pp. 346–353, 2010, doi: 10.1007/978-3-642-12079-4_43.
 - [30] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Bus. Horiz.*, vol. 53, no. 1, pp. 59–68, 2010, doi: 10.1016/j.bushor.2009.09.003.
 - [31] S. S. Singh, K. Singh, A. Kumar, H. K. Shakya, and B. Biswas, *A Survey on Information Diffusion Models in Social Networks*, vol. 956, no. January. Springer Singapore, 2019.
 - [32] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 599–608, 2010, doi: 10.1109/ICDM.2010.22.
 - [33] D. Agrawal, C. Budak, and A. El Abbadi, "Information diffusion in social networks," *Proc. VLDB Endow.*, vol. 4, no. 12, pp. 1512–1513, 2011, doi: 10.14778/3402755.3402811.
 - [34] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proc. R. Soc. London. Ser. A, Contain. Pap. a Math. Phys. character*, vol. 115, no. 772, pp. 700–721, 1927.
 - [35] "Understanding Information Cascades in Financial Markets."
<https://www.investopedia.com/articles/investing/052715/guide-understanding-information-cascades.asp> (accessed Jun. 13, 2021).
 - [36] D. Easley and J. Kleinberg, "Networks, crowds, and markets: Reasoning about a highly

- connected world,” *Significance*, vol. 9, no. 1, pp. 43–44, 2012.
- [37] M. Granovetter, “Threshold models of collective behavior,” *Am. J. Sociol.*, vol. 83, no. 6, pp. 1420–1443, 1978.
- [38] “(3) (PDF) Influencers characterization in a social network for viral marketing perspectives.” https://www.researchgate.net/publication/321900474_Influencers_characterization_in_a_social_network_for_viral_marketing_perspectives/figures?lo=1 (accessed Jun. 13, 2021).
- [39] S. Akrouf, L. Meriem, B. Yahia, and M. N. Eddine, “Social Network Analysis and Information Propagation: A Case Study Using Flickr and YouTube Networks,” *Int. J. Futur. Comput. Commun.*, no. January, pp. 246–252, 2013, doi: 10.7763/ijfcc.2013.v2.161.
- [40] C. Perez and R. Germon, “Graph creation and analysis for linking actors: Application to social data,” in *Automating Open Source Intelligence*, Elsevier, 2016, pp. 103–129.
- [41] “About.” <https://gephi.org/about/> (accessed Jun. 17, 2021).
- [42] “About Python™ | Python.org.” <https://www.python.org/about/> (accessed Jun. 18, 2021).

Abstract:

Social influence is the science that studies how people's ideas, attitudes and actions may be changed in combination with Social Network Analysis (SNA) where relationships and flows are important to map and measure. The problem of this research is influence and information diffusion. This analysis is based on the Linear Threshold Model and the Independent Cascade Model (ICM) using a dataset (network): the Facebook user's is based on its structure (size, type and diameter, components) and type of relationships. Network metrics and visualization were manipulated by Gephi as well. Finally, the findings of our experiments indicate that the selection of the starting nodes has a major effect on the diffusion process. and (LTM) model gives more nodes active than (ICM) model.

Key Word: Social Network Analysis (SNA), information diffusion, influence

Résumé :

L'influence sociale est la science qui étudie comment les idées, les attitudes et les actions des gens peuvent être modifiées en combinaison avec l'analyse des réseaux sociaux (SNA) où les relations et les flux sont importants pour cartographier et mesurer. Le problème de cette recherche est l'influence et la diffusion de l'information. Cette analyse est basée sur le modèle du seuil linéaire (LTM) et le modèle en cascade indépendant (ICM) à l'aide d'un ensemble de données (réseau) : l'utilisateur Facebook est basé sur sa structure (taille, type et diamètre, composants) et le type de relations. Les mesures et la visualisation du réseau ont également été manipulées par Gephi. Enfin, les résultats de nos expériences indiquent que la sélection des nœuds de départ a un effet majeur sur le processus de diffusion et le modèle (LTM) donne plus de nœuds actifs que le modèle (ICM).

Mot clé : Analyse de réseau social (SNA), diffusion de l'information, influence.

ملخص:

التأثير الاجتماعي هو العلم الذي يدرس كيف يمكن تغيير أفكار الناس ومواقفهم وأفعالهم بالاقتران مع تحليل الشبكة الاجتماعية (SNA) حيث العلاقات والتدفقات مهمة لرسم الخرائط والقياس. ومشكلة هذا البحث هي التأثير ونشر المعلومات. ويستند هذا التحليل إلى نموذج العتبة الخطية والنموذج التعاقبي المستقل باستخدام مجموعة بيانات (شبكة): ويستند مستخدم فيسبوك إلى بنيته (الحجم والنوع والقطر والمكونات) ونوع العلاقات. وتلاعب غيبي أيضا بمقاييس الشبكة والتصور. وأخيرا ، تشير نتائج تجاربنا إلى أن اختيار العقد الأولى له تأثير كبير على عملية الانتشار. ونموذج (نمور التاميل) يعطي المزيد من العقود النشطة مقارنة بنموذج (ICM).

الكلمة الأساسية: تحليل الشبكة الاجتماعية (SNA) نشر المعلومات ، التأثير .