Social network analysis : detection of influencers in fashion topics on Twitter

Auteur : Tridetti, Stéphane
Promoteur(s) : Ittoo, Ashwin
Faculté : HEC-Ecole de gestion de l'ULg

Diplôme : Master en ingénieur de gestion, à finalité spécialisée en Supply Chain Management and Business Analytics
Année académique : 2015-2016
URI/URL : http://hdl.handle.net/2268.2/1348

SOCIAL NETWORK ANALYSIS: DETECTION OF INFLUENCERS IN FASHION TOPICS ON TWITTER

Promoter:

Ashwin ITTOO

Readers:

Michael Schyns

Stephanie Aerts

Dissertation by :

Stephane TRIDETTI

For a Master in Business

Engineering specialized in

Supply Chain Management

Academic year 2015/2016

Acknowledgements

These last 8 months during which I wrote this Master thesis were a truly enriching experience in both terms of professional learning but also at a personal level. I would like to acknowledge the persons who have supported and help me throughout this period.

I would like to express my sincere gratitude to Professor Ashwin Ittoo for giving me the opportunity to work with him, for the continuous support, for his availability, patience, motivation, enthusiasm and immense knowledge. He really provided me a valuable support during these whole year.

Beside my advisor, I would also like to thank the rest of my jury: Professor Michael Schyns and Ph.D Stephanie Aerts for their availability and for taking the time to read this thesis.

I would also like to thank my friend, Ph.D student Sebastien de Bournonville, his help to solve the bugs during the implementation of the project and also for presenting me Latex, the tool I used to write this thesis.

Last but not least, I would like to thank my family for supporting me spiritually during my study, especially this project.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Social network sites have facilitated and stimulated the interaction and communication in real-time. Users are engaged by sharing their opinions, experiences, feelings on various topics such as news, politics, personalities, events or products. Due to the growth of online social networks, they became a rich and important source of information and with no surprise have generated interest of many companies.

Used correctly, it can accelerate their brand messages and contents. Fashion industry has probably the most potential in digital marketing solutions and has integrated them in their brand communications. In recent years the use of digital influencers became a new strategy in the development and in the management of marketing campaigns for leading brands and companies. Nowadays, they became brand ambassadors in communication of brand messages and help them to create involvement and brand awareness in a specific target audience through continuous brand activations. Fashion houses empower digital influencers to design beautiful and inspiring content including photographies or videos through social network sites and blogs.

Twitter is one of the most notable, potential and powerful social network sites to market businesses because it provides real-time data for business insiders. Unlike the other social network services, Twitter employs a social-networking model called "following" (Weng et al., 2010) where nobody needs permission to follow someone, called friend.

The interconnectedness between actors shows to businesses how it is important to find out who are the influencers and the experts in their target area in order to understand their requirements because they ensure a high level of interaction within their community, and therefore potential consumers (Booth et al., 2011). In this context, the successful implementation of this marketing strategy requires the identification of those members who are structurally well integrated into a social network. When identified, they will be useful in various corporate issues, especially in the word-of-month marketing such as influence public option (Katz, 1995), improve the adoption of innovation (Rogers, 1995) or brand awareness (Keller, 2003), and in viral marketing campaigns (Cha et al., 2010).

Given these findings, it looks appropriate to tackle this issue with the social network analysis that has already developed a variety of centrality measures to quantify the interconnectedness of actors and ranked users based on their influence on social network.

Therefore, this dissertation aims (1) to describe a current state of research in centrality measures on social networks in terms of interpretability, robustness and accuracy and current applications of these measures, (2) to propose a new method to collect data from Twitter with a given topic by using a friendship graph and avoiding problems that text-processing can procure, and (3) to provide a new empirical study on centrality measures in fashion industry which has not been treated yet in the literature.

This paper will be organised as follows: in the chapter 2, the representation of a social network as a graph will be parsed with terminologies, definitions and properties. Afterwards, the role of influencers in a marketing context will be defined. Moreover, general definition of centrality measures and formulations of degree, betweenness, closeness and eigenvector will be detailed in a context of undirected or directed and weighted or unweighted networks. The last part of chapter 2 provides the current state of research in centrality measures on social networks. The chapter 3 explains the methodology features to facilitate its reproduction. On the one hand, the identification of Twitter influencers in fashion topics will be addressed. After that, a step-by-step algorithm will be provided to gather Twitter data and to construct a friendship graph. On the other hand, algorithms are constructed with these specific centrality measures since the literature does not provide a step-by-step approach. To finish the chapter 3, the specific aspects of the implementation will be detailed. Finally, the last chapter presents the results and discusses them, including the limitations of the models. A final conclusion with my contributions to this research will end this dissertation.

Chapter 2

Literature Reviews

2.1 What is a Social Network?

The idea of *social network* emerged in the 1890s when German scientists carried out research on social groups. It is only in the 1930's, however, that many developments have appeared in the field of psychology, anthropology, sociology and mathematics (Scott, 2000) with currently thousands of applications.

The term social network is understood in this paper as "a set of socially relevant nodes connected by one or more relations" (Scott et al., 2011, p.12) that represents characteristics of friendships, advices, communications or other supports that occur among adherents of a social system (Valente, 1996).

Nodes, vertices or points are all synonyms and represent individual actors such as employees in a corporate work team, terrorists preparing an attack or high school students attending a prom, or collective actors such as political parties holding in a parliament or firms competing in an industry (Knoke, 2008). Links or edges denote the interconnectedness between actors : how one member is in relation with another member within his network.

Borgatti, Mehra, Brass & Labianca (2009) identified three types of relation that can exist in a social network: one based on similarities in which two nodes share the same attributes or variables such as demographic characteristics, locations or group memberships; a second one based on the social relations, characterized by the cognitive awareness or affection between two nodes; and the last one, the interaction relationship that refers to a continuous exchange of information between two actors such as speaking, chatting and so forth.

2.1.1 Social Network Modelling

This section provides graph-theoretical terminologies on basic concepts that I will use in the rest of the research. For more details, see Benzi & Klymko (2014), Knoke (2008) or Scott (2000).

In the mathematics side, a social network is represented as a *graph* $G = (V, E)$ by a set of n nodes (also called points or vertices) V with $|V| = n$ and a set of edges m where $E = \{(u, v)|u, v \in V\}$ (Benzi et al., 2014). In this dissertation, the notation (i, j) is also used in case of (u, v) in order to name the link from node i (respectively vertex u) to node j (respectively vertex v).

A graph G is *undirected* if "the edges are created by unordered pairs of vertices" (p.2). Each vertex v can be qualified by the number of edges incident to v in G, called *degree* σ_D. A *walk* of a length k in a undirected graph is "a set of nodes $v_1, v_2, ..., v_{k+1}$ such that for all $1 < l < k$, there is an edge between v_l and v_{l+1}" (p.2). A *path* is "a walk with no repeated nodes" (p.2). The *distance* between two vertices in the graph is "the number of edges in a shortest path (also called geodesic distance) connecting them" (p.2). However, the shortest path between two nodes is not unique, there may exist more than one. A graph G is considered *connected* if a path exists between all pairs of nodes. A *cycle* is "a path with at least distinct edges in which the source (the first) and the target (the last) are the same" (p.2).

A graph G can also be *directed* (also called *digraph*) if edges are formed in ordered pairs of vertices such that $(u, v) \in E \not\Leftrightarrow (v, u) \in E$. In a directed graph, a vertex has two types of degrees. The *indegree* σ_D^{in} of a node v is given by the number of edges point in v, while the *out-degree* σ_D^{out} of a node v is given by the number of edges point out v. The definitions of path and cycle are extended to directed graph. However, in a digraph, the *distance* $d(u, v)$ between two nodes u, v is defined as "the *length* of a shortest path from nodes u, v" (p.2). Compared to undirected graph, $d(u, v)$ is not necessarily identical with $d(v, u)$.

The *neighbourhood* $N(v)$ of a vertex v in a graph G "is the set of vertices adjacent to

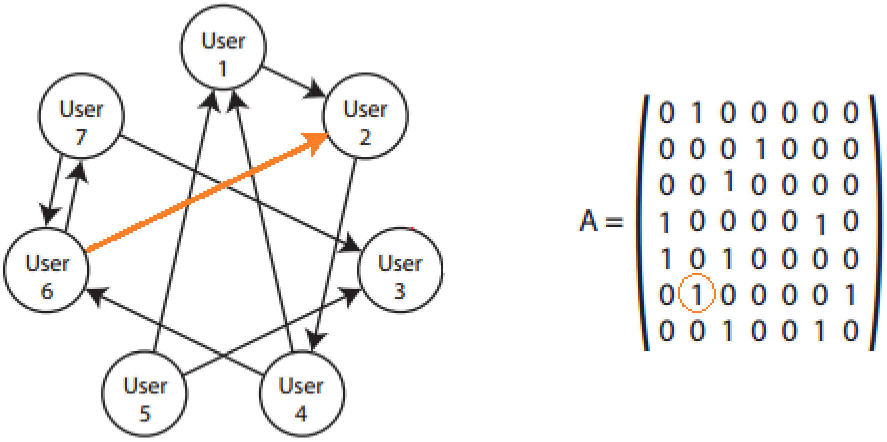
v" (p.2). The neighbourhood does not include v itself.

A graph G is a structured data that can be represented by a squared adjacency matrix $A = (a_{uv})$, where the size of the matrix is equal to the number of vertices in the graph and the entry (Benzi et al., 2014),

$$a_{uv} = \begin{cases} 1 & \text{if an edge } (u, v) \text{ exists in } G, \\ 0 & \text{otherwise} \end{cases}$$

The adjacency matrix is the starting point of the centrality computation. Some properties must be pointed out to understand their calculations. If the graph G is undirected, the adjacency matrix A =(a_{uv}) $\in$ {0;1}^{nxn} is binary and symmetric with all the diagonals equal to 0. Since A is symmetric, then A is reducible and diagonalisable (Bair, 2011). If the graph G is directed, its adjacency matrix A is also binary but not necessarily symmetric. If all the eigenvalues of the matrix A are distinct, then A is diagonalisable (Bair, 2011).

For example, considering a directed graph as shown in *Figure 2.1* and where V_G = {v_1,...,v_7} and E_G = {e_1,...,e_{10}}. The edge e_2= (v_6, v_2) means that a relation exists only from the user 6 to the user 2 but the reciprocity is not true. As seen before, the representation of the graph must be expressed in rows and columns. For instance, the entry a_{62} is 1 but the entry a_{26} is 0 because there is no link directed from user 2 to user 6.



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Figure 2.1: Example of graph structured data expressed in an adjacency matrix

A graph can also be weighted which means an edge e is associated with a real number $w(e)$ (or $w(uv)$), called weight. In social network, the value of $w(e)$ is considered as a non-negative integer relying on the "cost" of the edge e.

2.1.2 Network properties

Network properties in social network are relevant to understand the structure of the network but also how some models contrast to each other. For more information refer to Easley & Kleinberg (2010).

The *size* of the network refers to the number of nodes it contains. A graph is *dense* if "the number of links is close to the maximal number of links that the graph can contain" (p24). The measure of *reciprocity* relies on "the proportion of mutual connections in a digraph. This concept is commonly defined as the probability that the opposite counterpart of a directed link is also part of the graph" (Csardi et al., 2015, p.312). The *clustering coefficient* of a node v is the probability that "two randomly selected friends are friends with each other in a undirected graph. The value is between 0 (when none of the node's friends are friends with each other) and 1 (when all of the node's friends are friends with each other)" (p.49). The *homophily* of a graph relies on the principle that people tend to be similar to their friends and to share common characteristics that ease the creation of communication and relationship. In this thesis, the presence of homophily must be proved to valid the sample extracted, and afterwards, applying meaningful centrality measures on this sample.

2.1.3 Social Network Analysis

Since many years, mathematical and computational studies of complex networks[1] have been a central issue in many fields such as biology, sociology, finance, computer sciences, physics...

Social network analysis (SNA) is the branch of sociology where they map and quantify the interconnectedness between actors in a network through mathematics. Aims are to understand the structural relations and to explain both why they occur and what are their consequences (Scott, 2012). This interdisciplinary academic field has many advantages derived from the graph theory because it provides the vocabulary/concepts and properties

[1] A *complex network* is characterized for not being regular (connections have not clear and homogeneous characteristics). In this paper, complex network and network are used interchangeably.

that allow researchers to prove theorems about graph and allow the representation of social structure (Freeman, 1979 ; Wasserman, 1994).

In this case, terminologies from sociology and graph theory may be linked. In this thesis, network and graph terms will be used interchangeably. As said before, a matrix is relevant to represent network structures and at the end to explain their effects. The adjacency matrix is also called *sociomatrix* and can be represented by a *sociograms*, but it is not necessary to draw sociograms to use graph theoretical concepts and measures (Scott and al., 2011).

The guiding assumption of SNA is that the way used by the members of a group to communicate to each other affects some important characteristics of that group, e.g. efficiency when performing a task, moral satisfaction, leadership (Knoke, 2008; Scott, 2012). This must be compared with the part of social sciences assuming that actors take decisions and act without regard to the behaviour of other actors.

Studies on off-line social networks have shown interesting properties such as high clustering coefficient (Newman, 2004), "small word" or six degrees of separation effect (Milgram, 1967) or scale-free effect[2] (Castells, 2004). Moreover, scientists have demonstrated similar patterns on online social networks (e.g. see Leskovec (2008) for small world phenomenon or Bichler (2008) for scale-free effect). The important element considered from Milgram research (1967) is that the majority of members in a social network can be assumed as a single graph. On this dissertation, only connected networks are assumed because if a disconnected sub-group of nodes is considered, then the distance has an infinite value.

In the literature, online social networks are called *social network sites* or *web social networks* and are defined as follows: "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) 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." (Ellison, 2007, p.210)

With the emergence of social network sites, it is not surprising that the need of characterization or comprehension of interconnectedness between actors has been a central issue for industries and scientists. Web-based services such as Facebook, Twitter and Linkedin

[2] The scale-free effect has shown that some nodes can act as a hub within their network.

have influenced considerably the communication and interactions between members which hold valuable information for companies. Nowadays, digital influencers are part of the communication strategy in order to market products and to engage the community to know how well their campaigns, brands or products are being perceived.

This next section will address the role of influencers and explain why it is so strategical to identify them for the success of business decisions.

2.2 Influencer in a Marketing Context

Web-based services are new playground of searchers and marketers to maximize the word-of-mouth diffusion of a message with specific *influential persons* or *influencers* (Friedl et al., 2010). In the literature, there is no agreement on what is an influential person (Riquelme, 2015). Therefore, measuring the influence of a user in a social network is a conceptual issue that conducted authors to develop plenty of measures with measurement criteria that differ from the others.

Nonetheless, two types of influencer might be distinguished in a marketing context. The word influencer can be defined either as an individual who impacts the spread of information: people who influence people (Weinmann, 1994), or as an individual who exhibits some combinations of desirable attributes such as trustworthiness and expertise or network attributes (connectivity or centrality) (Keller, 2003).

The first group of influencers is also called in the literature opinion leaders, prestigious (Gayo-Avello, 2013), innovators (Cha et al., 2013), key-players (Borgatti, 2006) and spreaders (Kiss et al., 2008). The second group of the definition is illustrated in the literature as celebrities (Srinivasan, 2013), evangelists (Bigonha et al., 2010) or experts (Keller, 2003) such as a journalist in BBC or a professor at Harvard University. Those terms will be used interchangeably in this paper but respectively to their category.

Measuring and quantifying the prestige of an influencer is relevant for marketers and businesses because they may touch a large scale of audience with a very small marketing cost and fast delivery, especially through technology as the World Wide Web. (Kozinets, 2010 & Valente, 1995; Friedl et al., 2010).

The role of an influencer in the word-of-month marketing as communication strategy can be used in many fields: possibly influence public opinion in the flow of mass communication (Katz, 1995), help business in product development by gaining market shares (Bass, 2004), improve the adoption of innovation (Rogers, 1995) or improve the band awareness (Keller, 2003). The influencer's role is also helpful in viral marketing campaigns by integrating well-connected actors in order to attract the attention of the largest possible audience to brand products or campaigns (Cha & Haddadi, 2010; Kiss & Bichler, 2008).

2.3 Centrality in Social Networks

In order to identify influencers in a social network, it looks appropriate to use *centrality measures* (CM) that has been developed within the social network analysis.

2.3.1 What is Centrality?

Similarly to the definition of influencer in a marketing context, "there is certainly no unanimity on exactly what centrality is or on its conceptual foundations, and there is very little agreement on the proper procedure for its measurement" (Freeman, 1977, p.217). Due to the problem to know what centrality is, one can only determine what centrality does (Freeman, 1980). The main contributions of centrality measures are to measure and quantify the "structural features" of a single node in a graph (Freeman, 1979), especially to find actors with a central position in the graph. It determines how a node is "important" in the network and it provides a ranking with the most important nodes (Bonacich, 1987).

Different researches show that centrality measures have been relevant indicators of the social power of actors and their abilities to explain how well they are connected with the network (Bonacich, 1987; Knoke & Burt, 1983). They proved great values in the analysis and comprehension of influencers in a social network. For instance, Wasserman & Faust (1997) demonstrated that individuals with a *central position* in a network tend to be more active to maintain and manage their own contacts by trying to reduce the paths with other nodes in the network, which means increasing the number of direct links. Moreover, an influencer owns unique social advantage for acquiring information and resources (Cross et al., 2001) that implies a better control over information acquisition in order to share them.

In this thesis, a node which has a central position (also called *node-centrality*) and influencer have the same signification because only influence based on network attributes is considered.

Centrality measures are also beneficial to measure the popularity of a vertex. Members with a high degree of popularity have high access to others and have a large number of people who are willing to share information with them (Cross et al., 2001). However, high popularity and high influence are not well correlated which means that popularity does not imply influence and vice-versa (Romero et al, 2011).

Other methods are available to detect how a vertex is "important". Social capital (e.g. Borgatti, Jones & Everett, 1998; Burt, 1997) is another field which attempts to quantify and identify them. Nevertheless, the social capital research are more focused on the features of the network that contribute to the individual, compared to centrality where we want to know who is important for the network (Borgatti, 2006).

2.3.2 Common Measures of Centrality

As discussed before, plenty of measures of node centrality have been developed and it is impossible to present here all of them. However, there are four basic concepts that have been highly developed in the literature : degree, betweenness, closeness and eigenvector. Indeed, each concept generalizes different interpretations of a node centrality in a graph (Freeman, 1979) and most of other developed centrality measures are derived from them. For example, PageRank algorithm (Page et al., 1999) has the same conception as eigenvector centrality. This sub-section will provide the definition for these four basic concepts in undirected or digraphs and/or unweigthed or weighted graphs in a social network context. Moreover, a general formulation and the interpretation of these measures are shown and analysed.

In our graph structure data, all our networks are assumed as *static* structures. Indeed, a snapshot of the vertices and edges of a particular network for a moment in time is considered (Easley et al., 2010).

Degree Centrality

Provided by graph theory, *degree centrality* (DC) σ_D is defined as the number of contacts a node may have in the network (Nieminen, 1974 ; Freeman, 1978).

The centrality score for a vertex i is higher, the more contacts or edges a node i has and this vertex is considered as a hub (Freeman, 1978).

The formula can be represented, for an undirected and unweighted graph G, as follows (Opsahl et al., 2010):

$$\sigma_D(i) = \sum_{j}^{N} a_{ij} \tag{2.1}$$

where i is the focal node and j represents all others nodes, N is the total number of nodes and a is the adjacency matrix in which the entry a_{ij} is equal to 1 if node i is connected to node j. Otherwise it is equal to 0. Note that the adjacency matrix is not necessary for the computation here, one can just visually count the number of links attached to a node. However, this formula is impacted by the size of the network. In order to compare the obtained results with other graphs, normalisation should be applied on the calculation by dividing them by $n-1$ vertices. Normalisation provides a value between 0 and 1.

In a digraph, this measure is split in indegree centrality σ_D^{in} and outdegree centrality σ_D^{out}. Illustrated in figure 2.2, indegree centrality counts the number of edges directed to the node i, whereas outdegree centrality counts the number of edges directed from node i to others. Indegree is a good proxy of popularity, while outdegree represents the activity of a node or its gregariousness (Opsahl et al., 2010).

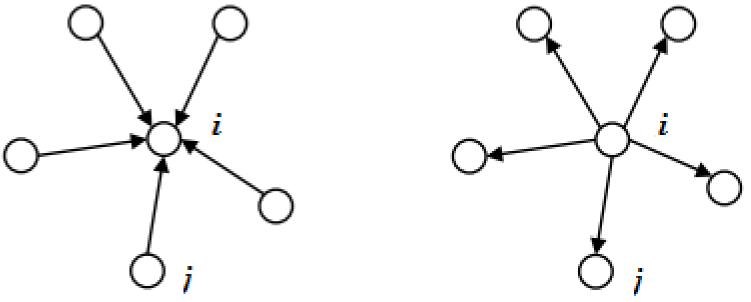


Figure 2.2: Example of indegree and outdegree CM

In a weighted graph and undirected network, the σ_D^w of a node i is simply the sum of weight w_{ij}, where w_{ij} is equal to 1 if a connection between node i and node j exists, otherwise is equal to 0 (Newman, 2004).

$$\sigma_D^w(i) = \sum_j^N w_{ij} \tag{2.2}$$

where $w_{ij} > 0$ if an edge exists between node i and node j, otherwise $w_{ij} = 0$.

Opsahl, Agneessens & Skvoretzc (2010) have investigated the computation of degree centrality in weighted and directed graph, unusual for the analysis of social networks in this thesis.

Closeness Centrality

Closeness centrality (CC) is another point of view to characterize the central position of a node in a graph. The idea of CC was proposed by Bavelas (1950), Beauchamp (1965) and Sabidussi (1966). It relies on how a node may control the communication in a network. Those authors proposed different interpretations of CC but the general definition can be understood as follows: a node is central if it is independent upon other as intermediaries (or re-layers) of a piece of information (Freeman, 1978).

However, Bavelas (1950) suggested that the most central node can spread a message throughout the entire network with the minimum time, while Sabidussi (1966) relied it to the node that spends the minimum time or cost to communicate with all others in the network. Beauchamp (1965) also defined it as the "optimum efficiency" in communication.

The independence of a node to others can be illustrated graphically with the figure 2.3. For example, the node 2 is directly connected with nodes 1, 3 and 4. If it wants to reach node 5, it must pass the message through node 4. We achieved that if node 2 wants to communicate with the complete network, node 2 depends only on one intermediary. However, node 1 needs node 2 to reach nodes 3 and 4 but also node 2 and node 4 to communicate with node 5. In order to reach all nodes in the graph, node 1 must depend three times on node 2 and once on node 4 (there are four intermediaries). In conclusion, node 2 has a greater centrality than node 1 because there are less intermediaries or less independent nodes to others because it is closer to all other nodes (Freeman, 1978).

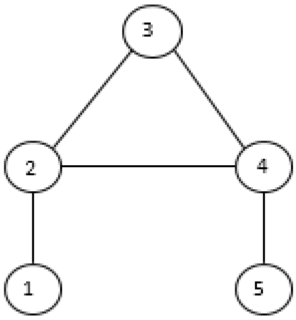


Figure 2.3: Graph with 5 vertices and 5 undirected edges (Freeman, 1979, p218)

Mathematically, the simplest computation of closeness centrality σ_C was proposed by Sabidussi (1966) by summing the geodesic (shortest paths) inverse distances from a source vertex to all other vertices in the network. The formulation, for both directed and undirected graphs, can be represented as follows (Opsahl et al., 2010):

$$\sigma_C = \frac{1}{\sum_j^n d_G(i,j)} \tag{2.3}$$

where $d_G(i,j)$ is the number of links in the geodesic distances from node i to node j. We achieved that σ_C increases when the distance to another node is reduced. Nevertheless, all nodes in the graph must be connected because the distance to one node to all others must be a finite distance.

However, considering the distances *from* or *to* all other nodes must be precised in a digraph. The distance *to* a node is a more significant measure of centrality, which means a node has little control over its links pointed to it (Freeman, 1978).

A breath-first search (BFS) algorithm should be used to identify the shortest paths in a larger binary network. Again, the value of σ_C is impacted by the size of the network. Therefore, we can not compare the value of σ_C for graphs of different sizes. Beauchamp (1965) solved the effect of network size by dividing σ_C by $(n-1)$ nodes present in the graph.

In a weighted network, the formulation can be generalized as follows (Opsahl et al., 2010):

$$\sigma_C^w = \frac{1}{\sum_j^n d_G^w(i,j)} \tag{2.4}$$

However, in order to find the geodesic distances in a weighted network, the Dijkstra's algorithm (Dijkstra, 1959) is the best suitable solution for this problem. Weighted typology is not really present in the literature of social network analysis.

Betweenness Centrality

Betweenness centrality (BC) was developed by Freeman (1978) who defined it as follows: "a point that falls on the communication paths between other points exhibits a potential for control of their communication. It is this potential for control that defines the centrality of these points" (Freeman, 1978, p 221). In fact, these nodes are called boundary-spanner and may act as a bridge between groups of nodes and may influence groups by withholding or distorting the transmission of information.

To explain the formulation, Freeman (1978) first defined the "partial betweenness" of a graph (p.222) in terms of probabilities. For example, in the figure 2.3, there are two geodesics linking node 1 with node 3 (via node 2 and via node 4 respectively). Node 2 and node 4 are both between node 1 and node 3 and neither can control transmission of the information.

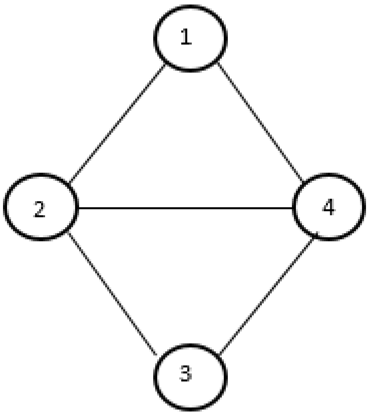


Figure 2.4: graph with 4 vertices and 5 edges (Freeman, 1979, p.223)

To generalize it, the probability of using a specific geodesic is $\frac{1}{g_{ij}}$, where g_{ij} is the number of geodesics linking node i and node j. The probability $b_{ij}(k)$, that a node k falls on a randomly selected geodesic connecting node i and node j, represents the potential of the node k for the control of information passing between node i and node j. Mathematically, it can be depicted as follows (Freeman, 1979):

$$b_{ij}(k) = \frac{g_{ij}(k)}{g_{ij}} \tag{2.5}$$

where $g_{ij}(k)$ is the number of geodesics linking node i to node j and containing node k. Indeed, it is an index of the degree to which the node i and the node j need the node k to communicate along the shortest paths linking them together.

The overall betweenness centrality $C_B(k)$ of a node can be found by the sum of its partial betweenness values for all unordered pairs of nodes where $i \neq j \neq k$ (Freeman, 1979):

$$C_B(k) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{g_{ij}(k)}{g_{ij}} \tag{2.6}$$

It becomes more complicated to count and locate geodesics with large networks. Matrix methods should be provided to carry out this task (provided after).

Again here, $C_B(k)$ depends on the network size. Freeman (1979) divided the formula by $\frac{n^2-3n+2}{2}$ which is the maximum value taken by $C_B(k)$, accomplished only by the central point in a star. Compared to CC, the formula can be applied on disconnected and connected graphs.

The obtained values are between 0 and 1. If $C_B(k) = 1$ node i and node j can not communicate passing through node k, which means that the position of node k is important by exercising a high control over i, j. The same reflection can be used to compute the BC in a digraph (White et al., 1994). However, we will use a BFS directed to find the number of geodesic distances.

Measuring BC in a weighted network arises complication but applications exist such as in the transmission of diseases that is more likely to happen from a person to another if they have frequent interactions (Valente, 1995): weights here are the frequencies. More generally, the weight can be understood as the cost of the edge in which a high value represents a weak or costly tie, while a low value represents a strong or cheap tie (Opsahl et al.,2010). Again here, difficulties are based on the selection of an attempt to identify shortest paths in a weighted network. For example, Newman (2001) implemented the famous algorithm presented by Dijkstra (1959) to find the least costly path.

Linked to the unweighted network formulation, it can be illustrated as follows (Opsahl et al., 2010):

$$\sigma^w_{BC}(i) = \frac{g^w_{jk}(i)}{g^w_{jk}} \tag{2.7}$$

which is a combination of the number of intermediary nodes and tie/link weights.

Finally, to generalize the formulation for a weighted and directed network, the reasoning is similar but with an additional constraint in the identification of the shortest paths: "a path from one node to another can only follow the direction of present links" (Opsahl, 2010, p.248).

Eigenvector Centrality

Bonacich (1972) proposed a new interpretation of node centrality by introducing *eigenvector centrality* (EC). This measure is established on the idea that a node is "important" if its neighbourhood is also important which means that it depends on the number of adjacent nodes and their value of centrality. Compared to DC, which takes into account the number of direct links, EC takes also into account the indirect contacts in the network.

First, we have to define the set of neighbours $N(i)$ of a node i. Graphically, in an undirected graph, this simply represents its set of neighbours, while for a digraph, this is the set of neighbours which point to i (indegree). Then, we can compute a node's eigenvector centrality $X(i, G)$ such as:

$$X(i,G) = \sum_{j \in N(i)} X(i,G) \tag{2.8}$$

However, if we consider the adjacency matrix A_{ij} as diagonalisable and stochastic[3], then the homogeneous linear system $AX = \lambda_i X$ will deliver the eigenvectors associated with each eigenvalue λ_i. We can also write this equation in its following canonic form:

$$\lambda x_i = \sum_{j=1}^{n} a_{ij} x_j \tag{2.9}$$

where λ is a constant to ensure the equations have a non-trivial solution. As A is diagonalisable, we can use a power method to solve the equation. Also, because A is a non-negative and irreducible matrix, A always has a positive eigenvalue associated with an eigenvector with only positive entries.

To illustrate this findings[4], consider a graph G of 5 vertices and its edges represented by its symmetric adjacency matrix $A^{5\times5}$. I consider the same adjacency matrix from the website given in the footnote.

[3] A matrix A is stochastic if and only if A is non negative and the sum of each element of each row or column is equal to 1.

[4] source : http://djjr-courses.wikidot.com/soc180:eigenvector-centrality

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

In order to give a starting random positive amount of influence to each node, we consider the degree centrality of each vertex represented by a vector 5 × 1 such as $X = (3, 2, 3, 3, 1)$ respectively. Then, we multiple the vector X by the matrix A and we obtain another vector 5 × 1 that represents the new amount of influence for each node. In other words, the multiplication reassigns to each vertex the sum of the values of its neighbour vertices by spreading out the degree centrality.

We repeat this process, multiplication of the new vector X with A, until reaching equilibrium. In order to know when this equilibrium is reached, we can simply divide each value x_i by its largest value contained in X. This step is called normalization. We stop the process when the value x is not changing anymore. Since we are just adding things up, values in the vector X keep getting bigger, but we reach a point where the share of the total at each node will remain stable. Note that for each multiplication by A, only the component values of the vector X change but not its size.

In fact, when multiplying a vector by just a number, called a scalar, we just multiply the elements by the scalar. This process is known as *scalar multiplication*. In our example, the equilibrium is reached after three iterations with maximum eigenvalue for nodes 1 and 3. So, $X = (1, 0.75, 1, 0.88, 0.33)$.

As additional information, Bonacich (1987) presented a variation of EC by adding two parameters (α, β) for controlling the impact of global and local factors. It can be written as:

$$C(\beta) = \alpha(I - \beta A)^{-1}A1 \tag{2.10}$$

where α is a normalization constant that weighs the importance of the node's degree, β illustrates how important the centrality of the neighbours is, A is the adjacency matrix, I is the identity matrix and 1 is a matrix of all ones.

The interpretation is simple. First, if $\beta = 0$, then $C(\beta)$ is reduced to indegree, while if $\alpha = 0$, then $C(\beta)$ is reduced to the previous EC computation. In social network, only

$\beta > 0$ are considered which means nodes have higher centrality when they have edges to central nodes. Generally, a high value of β is considered in order to catch the global network structure matters (your friends, your friends' of friends and so forth).

2.4 Current State of Research

Due to the lack of critical reviews on centrality measures in social networks, this section provides the current state of research on the four basic concepts described above. However, most of these studies are carried out on unweighted and undirected networks but I tried to consider also digraphs. The Appendix I summarizes the approaches in a table of research used to discuss centrality measures in social networks.

2.4.1 Interpretability

One of the most drawbacks of those centrality measures is the difference in results that lead to different interpretations.

Freeman (1979) compared the DC, BC and CC in all possible graphs (including cycle, chain, star or wheel) of 5 nodes in undirected networks, and then, compared the scores obtained. He concluded that these three measures assigned the maximum centrality scores to a star or a wheel and to the circle graph the minimum score. On these forms (circle and star) that have particular network structure, he concluded that these three measures are agreed to assign extremes. However, there are differences in results for intermediate forms. Indeed, he found out that those three concepts are based on three competing theories of how centrality may influence group processes. He advised to specify if we compute either centrality as control (BC), or centrality as activity (DC) or centrality as independence (CC).

Eigenvector centrality is part of DC conception and it is considered as a generalization of DC. However, we decided to consider them separately due to the difference in the used calculation process.

This point of view can be also understood with a correlation analysis : are there centrality measures correlated? If the answer was yes, it would be somewhat redundant.

Valente, Coronges, Lakon, & Costenbader (2008) have investigated empirically the correlation among centrality measures by taking data from 62 sociometric networks in a variety of settings. They found strong but also varied correlations among CM. They considered DC and CC in a digraph and undirected network and BC and EC as undirected one. The highest correlation score was between EC and DC because both are symmetrised graph and consider direct connections. Indeed, EC is more strongly correlated with both, the symmetrized and asymmetric versions of centrality measures because the matrix computation considers the global network structure. The lesser correlated score was between in-closeness and out-closeness which expresses that "the direction matter is more significant than the property measured by the algorithm" (p.18). In a digraph, when the reciprocity in the network is high, there is less distinction between asymmetric and symmetric computations.

Due to the strong assumptions on which centrality measures are based, other measures have been created in order to be closer to the real situation.

Previously discussed, the BC proposed by Freeman (1977) is based on geodesic paths. However, in many real situations, communication does not travel through geodesic paths only. Freeman, Borgatti & White (1991) proposed a measure based on all possible paths between a couple of points (called *flow betweenness*). This measure counts all paths that carry information when a maximum flow is pumped between each pair of vertices. Also, Newman (2003) proposed a measure based on random paths, called *K-path centrality.*

In the same order of idea, scientists have proposed different CC more realistic for all types of communication scenarii. Noh and Rieger (2004) introduced a random-walk version which "measures the speed with which randomly walking messages reach a vertex from elsewhere in the graph" (p.3). Or, *hierarchical closeness* (Tran, 2014) that assesses the most important node as the node that can reach the most of nodes by the shortest paths whereas CC consider the most important node in a digraph simply as the node that can reach to the other nodes by the shortest paths.

In order to understand the limit of EC in a social network context, one should consider the following example. In a school, a student will increase his popularity if he received information from students who are themselves popular. This student who receives a lot of information from others would be a better and more valuable source of information. This is how EC is understood. Moreover, a student in a school may have some popularity that depends on its external status characteristics. EC catches these exogenous sources of status or information. This is the reason why scientists have developed different centrality measures based on EC to bridge the gap. We can cite for example *alpha centrality* (Bonacich et al., 2001), *PageRank* (Page et al., 1999) or Katz's (1953) centrality measure.

2.4.2 Robustness and Accuracy of Centrality Measures

Assessing characteristics and robustness of centrality measures is important because incomplete data sometimes occurs during the collection of data process.

Bolland (1988) assessed the performance of the four basic concepts through an empirical study in real and simulated networks. He used a dataset from a network of influent relationships among 40 participants in an education program. After that, he simulated the addition of random links to a target vertex and a randomly selected set of nodes in the network. The robustness to a random error and the sensitivity to a systematic variation in the network were then assessed. This study shows that BC is unstable and sensitive with a random variation and a systematic variation respectively of the network structure, particularly at the network center. Moreover, scores fluctuated in an unpredictable way. CC a is robust centrality measure even under considerable variation because results are correlated with the original data. However, EC is the least sensitive to systematic variation and the least pronounced for boundary-spanners. DC is also robust under considerable errors but if the density of the network does not change.

Another contribution on the robustness is provided by Borgatti (2006) who compared the four centrality measures by examining four types of error (adding and removing nodes and edges) on a random digraph. This study shows that the four centrality measures are similar with respect to pattern and level of robustness because they have shown a high correlation score with the original data. However, BC performed slightly worse (less accurate) than the three other centrality measures because BC is more sensitive and more affected to changes in the network.

Costenbader & Valente (2003) drawn the same conclusions as Bolland (1998). They studied the stability and the robustness of the four centrality measures when networks are sampled facing inaccurate or incomplete network data from an undirected and unweighted social network. In addition, they took a sample, called sub-graph, of the overall network and studied the correlation between the results of the sub-graph and the overall network.

It turned out that the stability of EC is the preferred centrality measure when the network data are incomplete due to its ability to capture entire network structures. Moreover, they characterized DC has the highest stability to a change, especially for node in the center of the graph. Again, BC fared less successfully than the other measures.

2.4.3 Conceptual Studies

Some scientists carried out studies on the characteristics and underlying assumptions of centrality measures.

Friedl & Heidemann (2010) drawn 3 general properties of centrality measures and assessed the DC, BC, CC and EC to know if the centrality measures fit the properties or not:

- The first property relies on the increase of centrality scores when the distance between an actor and another one is reduced through an additional relationship in the network.
- The second property relies on the increase of centrality scores when the number of paths with the shortest length from a member to at least one other increases through an additional relationship.
- The last property relies on the stability of the ranking between nodes when we add a new relationship between two actors.

DC does not meet properties 1 and 2 because the major disadvantage of DC is that it does not take into account indirect contacts. Based on the definition of CC, property 1 is satisfied. However, BC and EC do not meet any properties. EC formula is harder to interpret and is less comprehensive than the others centrality measures.

In the same study, Boland (1988) made two assumptions about the nature of the network flow: the first one concerning the deterioration of resources over distance and time and the second one related to the paths through which information are able to flow. He computed the deterioration rate as a function $s = d^{-x}$, where s represents the strength of resources after it has travelled distance d. If x increases, then the deterioration of resources decreases. He concluded that DC has no particular merit in its ability to identify central nodes because its function has an immediate deterioration. BC inflates the centrality to boundary-spanners and attenuates it for unimportant persons in the network. He suggested the usage of BC with networks containing a complex sub-group structure. He also suggested EC as a good predictor of change in terms of ability to attribute centrality to an individual who is close to acquire a central position. Indeed, it is due to its lack of decay function because EC is a non-geodesics acknowledgement.

2.4.4 Current Applications of Centrality Measures

Research assessed the suitability of centrality measures through different applications. The provided interpretations will be relevant afterwards for the discussion part. Logically and intuitively, off-line social networks consider only undirected graph. However, others techniques than centrality measures are used to characterize or predict the influence of a user (for example, refer to Baskshy et al. (2011) for decision tree).

Addressed previously, Freeman, Roeder & Mulholland (1979) re-examined their findings and evaluated the suitability of those three concepts (DC, BC and CC) to identify key-players on group communication and group problem solving. They used statistics to assess some patterns and variables of the network and then compared those results with the centrality scores. They concluded that BC is suited to find leaders in a group due to its potential for controlling communication. DC, which measures the activity, turned out to be important in understanding group performance (e.g. organizational suggestions, requests for information, answers, ...).

Hossain, Chung & Murshed (2007) studied the structured network position by extracting relational data of actors using a mobile phone and their abilities to disseminate information through the network. Again, they used different centrality measures such as DC, BC, CC and EC to find influencers. They concluded that actors with a high DC score are in a better position to disseminate information to others because they have many ties which means they can access and call more resources of the network. The BC scores had similarities with those from the DC results : these members are considered to have high influence because a lot of other members depend on their connections to communicate with others. Boundary-spanners have a great position for passing information to other groups in the network. Moreover, the results in CC scores had very low standard deviation compared to the mean. It seems that actors are quite close to each other which is difficult to interpret data.

Lee, Cotte & Noseworthy (2010), based on Burt's research (1999 and 2004), considered DC and BC as relevant indicators for the influence of individual customers on the behaviour of the entire customer base. First, they assessed the influence of the actors by ranking them regarding customers' self-assessment and the assessment assigned by them on other clients. Then, they compared it with the centrality scores of BC and DC. They associated BC with opinion leaders because individuals have the ability and privilege to share or manipulate information and integrate different clusters of people and subgroups with the network. They also pointed out that a customer with a high DC score relies on a high number of ties giving access to more diverse information about the network.

Other scientists (Yan & Ding (2009); Mutschke (2003)) applied centrality measures (CC, BC and DC) in co-authorship networks. Yan (2009) used co-authorship data from the field of library and information science from 1988 to 2007, while Mutschke employed data of the digital libraries research over a period of fifteen years. They reached the same conclusions by interpreting the sense of centrality measures with the data. BC performed best among the four centrality measures to show author's important to other authors' virtual communication. However, authors involved in different disciplines would have a high betweenness score while their role in this field may not be significant. Also, CC is relevant to understand the network structure and to quantify the author's position and its distance with others. However, CC does not measure the academic impact of a co-author: a co-author writing an article with an author having a high closeness score would also have a high closeness score. Based on these results, it is not surprising that scientists developed new centrality measures such as AuthorRank (Lui et al., 2005), for this particular application.

Bichler (2008) tried to find the best centrality measure that identifies influencers who achieve a maximum dissemination of a marketing message. They based the experiment on call data from a Telecom company, and then, they compared the performance of different centrality measures. The empirical study showed that, if the set of initial customers or nodes is selected regarding outdegree measures, the number of reached customers is significantly greater than when choosing the initial set of customers with other centrality measures across different usages.

2.4.5 Related Work on Twitter

A large number of empirical studies have addressed the matter of influencers and the diffusion process on Twitter by using different metrics that can be related to centrality measures. In the context of Twitter, only digraphs are considered.

The first issue on social network sites like Twitter is to determine the link structure of the network. Two types of network exist:

- On the one hand, the *follower/following network* or *friendship network* where a relation between two nodes exists if at least one follows the other one.

- On the other hand, the *interaction network* where a directed link between node i and node j exists if node i has cited (i.e. via mention, retweet or replay) node j.

The study performed by Cha, Haddadi, Benevenuto & Gummadi (2010) defined three measures on "influence in Twitter" : indegree, retweets and mentions across different topics and time. These three measures are a kind of centrality measures. The *indegree influence* of a user is defined as the number of followers he has, while the *retweet influence* relies on the number of retweets a user has, indicating its ability to generate content. The *mention influence* measures the number of mentions a user has, indicating its ability to interact with others in a conversation. They concluded that users with a high number of followers span a lot of celebrities or news sources. Retweets can be used to reinforce a message (e.g. when a group of users repeats the same message, the probability of adoption of the idea increases). Not surprising, mentioned users were mostly celebrities.

Compared to the presented applications in an undirected network, indegree centrality represents here user's popularity (Romero et al., 2011), but is not related to any notion of influence such as engaging audience, i.e. retweets and mentions. Moreover, outdegree is also difficult to interpret because the reciprocity in following does not exist and a user can follow everyone without asking permission. In this context, the passivity of an actor should also be considered (Romero et al., 2011).

Bigonha, Cardoso, Moro, Almeida & Goncalves survey (2010) detected evangelists and detractors on the brewery sector in Brazil on Twitter. They qualified influencers based on the popularity (sentiment analysis), on the network position (DC, BC and EC) and on other metrics to determine the quality of tweets. They used text-processing for topic-detection, and then, used the three centrality measures regarding the graph typology. In an interaction network, actors with a high betweenness score have an important role to disseminate information within the network, while actors have a high eigenvector value if they receive responses from many users or by assuming their connection are also important, from user with also a high score value. Moreover, indegree can also be viewed as an indicator of the number of times a member is cited or retweeted. In friendship networks, EC relies on important users who have many users that point to them or many other important users that point to them.

However, this last study shows how dealing with text-processing is not really efficient. For instance, the effort needed to train and construct the model is time-consuming and costly. Here, they trained their data manually. Moreover, manual classification is not a trivial task and led to the difficulty to clearly distinguish between a neutral and a positive tweet. In addition, as there is no benchmark for influential users' detection, this study shows also difficulties to build such a test collection.

Moreover, dealing with text-processing generates also problems in data extraction because the Twitter API[5] does not allow historical extractions more than one week. Secondly, the tweets are not provided in a uniform language and other metrics should be provided to assess the quality or the frequency of tweets in order to determine the influence of a user.

More generally, difficulties also appeared to control the size of the network that led to difficulties to represent and interpret the data. Also, the network is not ensured to be connected resulting to adapt the centrality calculations. In this thesis, we extracted a follower/following network in order to avoid all these issues and we ensured that the network was connected with a controlled size.

However, a study made by Weng, Lim, Jiang & He (2013) defined a method for topic-sensitive influential user detection in a follower/following network. The method considers a PageRank-like metrics to calculate the user influence of a sample of 1000 Twitter users based in Singapore. Some interesting findings are given in the analysis of the sample; it shows a reciprocal social network structure which means mutual following relationship, and also homophily implying that a user choose seriously the friends to follow because it is interested by their contents.

[5] *An API (Application Programming Interface) is a set of functions, protocols and tools that are used to build an application, or to facilitate the communication with services (Riquelme, 2015).*

However, previous studies (Cha et al., 2010; Moro et al., 2010) rejected the use of BC, CC and DC on friendship graphs because all users that receive the information are not necessarily interested by its content.

Contributions Marketers, planners or others agents who are interested in harnessing worth-of-mouth diffusion could find benefits from this study because it contributes:

- to provide a systematic review on the four basics centrality measures in unweighted, directed or undirected networks,
- to provide a quicker method to collect Twitter data with a given topic by using a friendship graph and with avoiding the problems coming from the text-processing,
- and, to bridge the gap by applying and analysing centrality measures in fashion topics.

Chapter 3

Methodology

3.1 Twitter

Twitter is a microblogging service with 320 millions of users that permits members to post messages (called *tweets*) of maximum 140 characters (Kwak, Lee, Park & Moon, (2010)). Being a follower on Twitter means that a user can automatically receive all of the messages posted by the followed account(s). However, we used in this paper *followers* as the number of users pointing to a member or a node (indegree) and *friends* as the outdegree relationships (links from a node to others). Followers can see all tweets, friends or other personal information.

Compared to other social network sites, the relation of following or being followed does not require reciprocity. A user can follow any other user, but there is no obligation for following back. Indeed, the social structure of Twitter can be viewed as a "pyramidal" category: some influence accounts, such as a journalist on BBC or a movie star, have millions of followers without following them back, while Facebook has a "circular" category where friendships are reciprocal (Feng, 2011). Users can also interact via text by using *retweets* with a "@" followed by the appropriate user name or via a "#" followed by an expression representing a *hashtag.*

In the growing body of literature on social network analysis, most of the studies and research have been published on Twitter data set. The advantages of finding influencers on this platform have multi-fold. First, Twitter provides real-time web data that gives a timely update of the thoughts of influential accounts. Secondly, since Twitter is used as a marketing platform (Milstein, 2008), targeting those influencers will increase the efficiency of a marketing campaign (e.g. a fashion manufactory can integrate influential users in make-up topics to potentially influence more people).

The main advantage of using Twitter gives the access to a huge amount of data and also an easier way to exact them. Twitter also offers an Application Programming Interface (API) that enables to obtain real-time access to personal information on users or tweets in a sampled and filtered form. However, personal information on Facebook users is difficult to obtain due to the protection established by the platform. Moreover, Facebook API allows only fetching users that authorized Facebook application; this leads to a poor data extraction and the impossibility to map a relevant network.

Likewise, other advantages, compared to others social network sites, are the huge amount of available documentation on the API with concrete examples available, the fact that Twitter API provides a standardized library across languages and last but not least an easier OAuth connection to the API. In 2011, Facebook won the title of the "worst API" in a survey of 1000 developers because a lot of bugs, poor documentation, slow response times, and others mentioned issues[1].

3.2 Overview of the Technique

At the beginning we have to know why it is interesting to study fashion topics. Recent years have shown that fashion digital influencers played an important role in brand communications. This evolution can be proved by simply looking at fashion-shows where bloggers are more and more invited. Being a fashion blogger became a full-time job where some of them have developed a high cash generation strategy. For instance, the founder of the blog *The Blonde Salad*, Chiara Ferragni, generated $7 millions in 2015[2]. However, we do not become an influential person in one day (Weng et al., 2011). They have gained credibility through concrete efforts and personal involvements to their community and can now reach a lot of people through their blogs and their social network sites.

How can we identify influential users on Twitter involved in the fashion domain? The methodology followed in this thesis will be separated into two parts. On the one hand,

[1] Source: http://techcrunch.com/2011/08/11/facebook-wins-worst-api-in-developer-survey/

[2] Source: http://edition.cnn.com/2014/10/01/world/europe/bloggers-six-figure-salaries/

we are going to describe step-by-step the extraction method used to map the network on users interested in fashion. In this process, (1) a list of 10 bloggers related to fashion topics is selected. Then, (2) we consider all the most commonly friends from our initial list, and select the 20 most commonly friends. After that, we gather all the friends in this new list of 20 fashion influencers, and again we add the 20 most commonly friends to our list (iteration 2). We repeat the process until reaching a list of 100 accounts after 5 iterations. Afterwards, (3) the friendship network is constructed and is translated into a binary adjacency matrix.

On the other hand, the different centrality measure algorithms (4) will be parsed and applied on this sample. The figure 3.1 provides an overview of the technique.

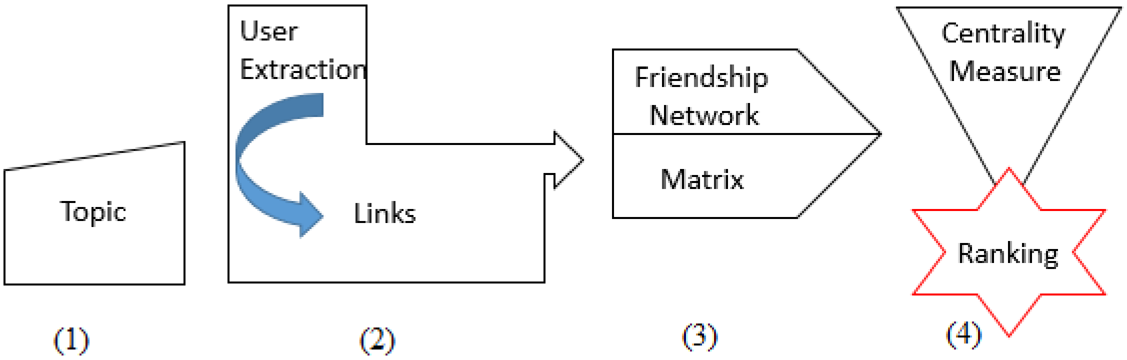


Figure 3.1: Overview of the technique

3.2.1 Complete Network Data

In order to collect our data, which means mapping the network, first we have to extract nodes and then to build the relationship between them. Some important assumptions should be made before.

First, only static and connected network is considered, reasons were detailed in chapter 2. Secondly, I considered a friendship network structure rather than one based on interactions such as mentions or retweets. Reasons are multiple. Because the thesis should be done in six months, this structure is the easiest way to extract Twitter data and avoiding text-processing problems described in the previous chapter. However, some authors (Weng, Lim, Jiang & He (2013); Cha, Haddadi, Benevenuto & Gummadi (2010)) rejected the application of centrality measures on friendship typology because it does not ensure that a follower is interested by the content published by a friend. To solve this problem, a strong assumption is made at the beginning of the extraction method: friends from an influence account for a specific topic should also be influent and should talk about the same topic.

3.2.2 Data Gathering

The following algorithm[3] is divided in different steps: steps 1 to 4 for the extraction of the needed nodes and the step 5 for the construction of the relations between users. We decided to limit our network size to 100 nodes because it is high enough to apply centrality measures on it. Also, dealing with more nodes should lead to difficulties in the comprehension and the analysis.

A step-by-step approach is provided here.

Step 1: Identify a list of 10 influential fashion bloggers

First, a list of 10 fashion bloggers is identified subjectively and manually. These influencers have already gained and maintained, through concreted efforts and personal involvements, the credibility and expertise in fashion advices. We want to avoid celebrities in this case because they would have others interests than fashion and they have probably gained credibility for others reasons.

To be as objective as possible, this list was computed by considering fashion bloggers in 2015 established by the magazine Fashionista [4]. They computed the ranking by taking into account different parameters such as the audience of each blogger on social network sites (Facebook, Twitter, Instagram, Youtube and Pinterest) and the personal website, the impact on brand extensions, and they asked to industry insiders who is the most efficient in selling products via affiliate links. The initial list is available on Appendix II.

The step 1 is the only one where a subjective way was used but is also an important one. It is critical to give a relevant first sample to the algorithm. This sample may be replaced by another one based on the experience in fashion.

[3] Source : http://www.captaindatascience.com/. This algorithm is based on this blog that deals with data management issues

[4] source : http://fashionista.com/2015/02/most-influential-style-bloggers-2015

Step 2: Extract all the friend accounts followed by users in the list

An argument in favour of outdegree selection of nodes in a follower/following networks is that Twitter has a pyramidal configuration, which not implies the reciprocity between friendships (Feng et al., 2011). In this context, we observed that experts, that are supposed to be influent (e.g. Obama, BBC ...), have a low number of friends but a high number of followers, reducing so the number of spanners or fake accounts (Ghosh et al., 2012).

For this second step, all friend IDs followed by the users from the initial list are extracted and are stocked in a database. Note that each element in the list will be referenced by an index i, e.g. the first ID in the list will receive the index 1 which will represent afterwards the node 1 in the network. Each list will receive this indexation feature.

In our list, Chiara Ferragni has only 153 friends, Aimee Song (367), Wendy Nguyen (378), Kristina Bazan (492), Zanita Whittington (863), Rumy Neely (304), Nicole Warne (829), Blair Eadie (313), Julia Engel (1145), Nicolette Mason (2048). Indeed, at the beginning, we exacted 10 different lists of Friend IDs from the initial list that we concatenated into one single resulting list, called list 1, of $153+367+378+492+863+304+829+131+1145+2048 = 6892$ IDs. However, I decided not to take into account the IDs of users in list A because if there are considered as influencers, they would appear in the next steps.

Step 3: Identity the 20 most common friends

We selected only the 20 most common friends followed by the initial list in a new list 1 which means that we considered the maximum occurrence that an ID can appear in the obtained list 1 at the end of the step 2 (e.g. if an ID appears 10 times, it will be ranked in the first position in the list 1, as the maximum occurrence an ID should appear is 10 here). For example, the Appendix III provides the IDs collected here.

However, in order to generalize the algorithm one should named list i for each $i = 0, ..., 5$. This step selects the 20 most common friends followed by the obtained users at the end of the step 2.

Step 4: Go back to step 2 and extract information on the 20 new IDs

Based on the new list i obtained at the end of the step 3, we repeated the step 2 in order to obtain a new list $i+1$ of 20 unique IDs. However, only unique ID is considered through iterations. For example, if list 3 has the same ID as one on all the previous lists (list 1 and list 2), we must remove the common ID in list 3 and must integrate the 21th ID because the size of each list i must be equal to 20 at the end.

We have to repeat the process until reaching 5 lists. The concatenation of these 5 lists, one should name it list X for a better comprehension, represents 100 unique IDs relying on the 100 nodes of the network.

Step 5: Build an adjacency matrix 100×100 with the obtained list X

Edges between nodes must be constructed in order to reach a follower/following network by a squared adjacency matrix $A = (a_{ij}) \in \{0; 1\}^{100 \times 100}$. Note that the size of our matrix is determined by the size of list X. Consider X the list of our 100 nodes $i = 1...100$ and its transpose Y with element $j = 1...100$. For each element i contained in the list X, friend IDs are extracted and compared with the elements in the list Y. If one element i is identical to one element j in the list Y, the entry in the adjacency matrix a_{ij} will receive a value equal to 1. Otherwise, the arbitrary value will be equal to 0. We will do that for each element in X and each line in the matrix. After that, a graph data structure of 100 nodes is obtained.

3.2.3 Centrality Measures to Rank Influencers

Centrality measures will be now used to quantify and rank users importance. Only directed and unweighted links are considered. The computation of DC will follow the computation described in chapter 2.

Closeness Centrality

In a first step, we have to define how shortest paths distances are identified. Indeed, we try to find the smallest number of edges that must be traversed in order to get to every vertex in the graph. We use a breadth-first search (BFS) algorithm because it is the most efficient way to calculate distances for a large network dataset (Easley et al., 2010). The algorithm starts from a source s (an arbitrary node in the graph) and explores the neighboring nodes, before moving to the next level neighbours. The particularity of this algorithm is that it pushes each reachable vertex onto the queue and considers each outgoing edge from it once.

A step-by-step algorithm (Easley et al, 2010, p.34) to trace out the distances in the global friendship network is provided:

"**Step 1:** we state all of your current friends to be at distance 1

Step 2: Then, we find all their friends (direct links and not counting persons who are already friends of yours), and state these to be at distance 2.

Step 3: we find all their friends (again, not counting persons already discovered at distances 1 and 2) and announce these to be at distance 3.

Step 4: Continuing in this way, we search in successive layers, each one expressing the next distance out. Each new layer is constructed from all those nodes that have not already been discovered in earlier layers, and that have an edge to some nodes in the previous layer."

Indeed, BFS assigns two values to each node i: a *distance* which gives the minimum number of edges in any path from the source s to node i, and a list of predecessors of node i along some geodesics from the source.

The minimum number of edges in any path from a node i and j can be generalized as follows: $d(i,j) = min(x_{ik} + ... + x_{kj})$, where k denotes the intermediary nodes on paths between node i and j (Opsahl et al., 2010). For example, if the two nodes are not connected together, but are connected to the same other node, the shortest distance between them would be 2.

The computation of the CC of a node i is simply obtained by inverting the sum of all

$d(i,j)$ from node i to all others nodes. Then, a normalisation must be applied to consider the size of the network by dividing the formulation by $(n-1)$.

Betweenness Centrality

Brandes (2001) provided the faster algorithm to give the exact centrality value of each vertex. Let denote graph $G=(V,E)$ with a node source s and a target t, σ_{st} is defined as the number of shortest paths between s and t, and $\sigma_s(v)$ the number of those shortest paths that pass through v (p.4). The Brandes's algorithm is composed of three steps.

Step 1: Compute the shortest paths, for each $s \in V$

By using a breath-first searching (BFS) for digraph, we find the shortest paths σ_{sv}, one for each $s \in V$. The BFS also provided the predecessor set $P_s(v)$ of a vertex v on shortest paths from s and the σ_{sv} values.

Step 2: For every $s \in V$, compute the dependencies $\delta_s(v)$ for all other $v \in V$

Brandes (2001) introduced the *pair-dependency* of s and t on v as the fraction of all the shortest paths from s to t through v over those from s to t.

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

He observed that "if a node v is a predecessor of a node w in a shortest path starting in s, then, v is a predecessor also in any other shortest path starting from s and passing through w" (Brandes, 2001, p. 4). Given this observation, he rewrote the pair-dependency formula as a recursive:

$$\delta_s(v) = \sum_{w:v\in P_s(w)} \frac{\sigma_{sv}}{\sigma_{sw}}(1+\delta_s(w))$$

where $P_s(w)$ is the set of direct predecessors of a certain node w in the shortest paths from s to w, thanks to the BFS rooted in a directed acyclic graph.

When we take a node w, its all dependency is calculated because they will have no future nodes further than w. The proportion of shortest paths from s to w passing through

v is given by $\frac{\sigma_{sv}}{\sigma_{sw}}$ which can be translated as the degree of dependency the vertices s and w on v to remain connected. Moreover, the vertex v is not only in the shortest path between s and w, it is also in all the shortest paths in which w is involved. We can explain the factor $1 + \delta_s(w)$ as follows: the factor 1 represents the dependency of v of the pair (s, w), and $\delta_s(w)$ relies on the dependency of v of every pair (s, "any vertex beyond w").

Step 3: Sum all dependency values

To calculate the centrality value of a node v, the sum of all dependencies should be done (Brandes, 2001, p.7):

$$C_B(v) = \sum_{s \neq v \in V} \delta_s(v)$$

Eigenvector Centrality

At the beginning, recall the intuitive process to calculate eigenvector centrality. First, we give to every node in the graph a starting random positive amount of influence. Then, each node splits its influence uniformly to its outdegree neighbours, receiving from its indegree neighbours in kind. The process ends when the system has reached a steady state which means until every node is giving out as much as it is taking in. At equilibrium, the amount of influence is represented by the eigenvector centrality.

To generalize the algorithm given by Bonacich (1972) (based on a the power algorithm), one should consider the adjacency matrix $A^{n \times n}$ that represents the binary relationship between actors.

Step 1: Construct A as non-negative adjacency matrix

Suppose a user P_j has l_j links. If one of those links is directed to user P_i, then P_j will pass on $\frac{1}{l_j}$ its importance to P_i. The importance taking of P_i is then the sum of the contributions made by all the users linked to it. So, we construct the matrix as follows:

$$a_{i,j} = \begin{cases} \frac{1}{l_j} & \text{if } l_j \text{ exists from } P_i \text{ to } P_j \\ 0 & otherwise \end{cases}$$

Each element a_{ij} is between 0 and 1 which defined A as a n-squared non-negative matrix. In this case, the sum of each elements of each line and each column are between 0 and 1 which implies that the module of each eigenvalues of A is also between 0 and 1. Moreover, the sum of each line is equal to 1, then 1 is eigenvalue of A associated with an eigenvector with all elements equal to 1 (Bair, 2011).

Step 2: Power method

The power method is an iterative way that repeatedly refines the estimation of the eigenvalues of a matrix, using the Rayleigh quotient of a non-zero vector (Parlett, 1974). One should consider $A^{n \times n}$. If x is an eigenvector of A with the associated eigenvalue λ, then there exist n linearly independent eigenvectors x_i $(i = 1...n)$ with $Ax_i = \lambda x_i$, and in general, $A^k x_i = \lambda^k x_i$ for all k. The observation is based on Perron-Frobenius theorem (see Bair (2011) for definitions and demonstrations) and represents the foundation of the power method.

Consider the set of $\{x_i\}$ of unit eigenvectors of A and its associated eigenvalues $\{\lambda_i\}$ such that $1 = \lambda_1 > |\lambda_2| \geq ... \geq |\lambda_n|$ (thanks to Perron-Frobenius theorem again). One should consider $v^{(0)}$ as an approximation to eigenvector of A with $||v^{(0)}|| = 1$. Then, a linear combination of eigenvectors of A from $v^{(0)}$ can be written; for some $c_1, ...c_n$ real value:

$$v^{(0)} = c_1 x_1 + ... + c_n x_n$$

and assume here that c_1 is not equal to 0.

Now

$$Av^{(0)} = c_1 \lambda_1 x_1 + x_2 \lambda_2 x_2 + ... + x_n \lambda_n x_n$$

and so

$$A^k v^{(0)} = c_1 \lambda_1^k x_1 + c_2 \lambda_2^k x_2 + ... + c_n \lambda_n^k x_n$$

$$A^k v^{(0)} = \lambda_n^k (c_1 x_1 + c_2 (\frac{\lambda_2}{\lambda_1})^k x_2 + ... + c_n (\frac{\lambda_n}{\lambda_1})^k x_n)$$

"Because the eigenvalues are real, distinct and ordered by decreasing magnitude, then for all $i = 2, ..., n$" (Panju, 2012, p.12)

$$\lim_{k \to +\infty} (\frac{\lambda_i}{\lambda_1})^k = 0$$

So, as k increases, $A^k v^{(0)}$ approaches $c_1 \lambda_1^k x_1$ and thus for large value of k,

$$x_1 \approx \frac{A^k v^{(0)}}{||A^k v^{(0)}||}$$

To sum up the process, the power method can be started as follows (Panju, 2011, p.12):

"Select a starting vector $v^{(0)}$ with $||v^{(0)}|| = 1$

For each k=1,2...

let $w = Av^{(k-1)}$

let $v^{(k)} = \frac{w}{||w||}$."

After each iteration, $v^{(k)}$ becomes increasingly closer to the eigenvector x_1 and the algorithm may be ended at any point with a great approximation to the eigenvector.

3.3 Implementation

3.3.1 R (Programming language)

We decided to implement the project with R. The advantages of this programming language are the following:

- R is free but also open source program. We can resolve bugs by yourself without waiting the vendor to fix them or add features in the future release.
- R supports both procedural and object-oriented programming where we can create your own objects, functions and packages and library.
- R is a relevant tool for statistics, data analysis and machine learning.
- R supports matrix arithmetics and others data structure such as vectors, lists, matrices, arrays and data frames [5].

[5] A data frame can be considered as a structured data which provide set of lists in the same size.

- R also provides crucial packages for our study such as `twitteR` that deals with the Twitter API and `igraph` that provides relevant functions to analyse the structure properties and compute centrality measures.

3.3.2 Data Extraction with TwitteR

Obviously, to use Twitter API we have to create a user account but we also need to obtain our keys and tokens from the Twitter Application Management [6]. On this page, we have to click on "Create New App" button to get access to our information.

As TwitteR (Gentry, 2011) is an external package to the basic R, we have first to install the package with the command `install.packages("twitteR")`. The TwitteR package needs to be loaded in order to gain access to and to use the various provided functions. We must load it by using the command `library(twitteR)`.

Then, we have to make the connection to the Twitter API with `setup_twitter_oauth()`. This function requires four arguments that represent our keys and tokens provided by the Twitter Application Management. Since they have long characters, we stored them in local variables and then call in the function.

Iteration 1

In the first iteration, we have to extract all the friend IDs of the given list of 10 influencers in the fashion industry. At the beginning, we concatenated, with the method `c()`, the 10 user names and stored them in a variable called `user_list`, as char type [7]. The major issue was to reduce the API call load. We used the function $lookupUsers()$ with two arguments : `user_list and includeNA=FALSE`. We stored this function in a variable called `user_accounts`. The lookupusers function will return a list of user objects ordered according to `user_list`. The argument `includeNA=FALSE` relies on that any non-existing users will be dropped from the list.

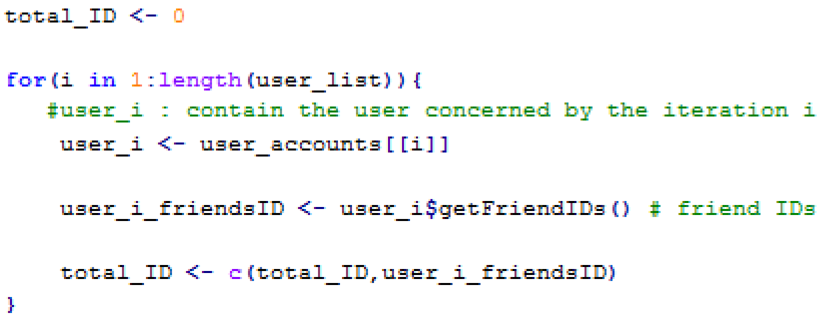
After that, we used a loop in order to extract the information of the user i. For each $i = 1...10$, we first stored in a variable, called `user_i`, the user object i contained in a list stored in `user_accounts`. The particularity of R is that the first element in a list starts

[6] On this page : https://apps.twitter.com/

[7] char = character type of data

at one and not zero compared to other programming languages. Then, we extracted the friend IDs for a given user name i by the request `user_i$getFriendIDs()`. It will return 10 lists i of friend IDs.

Note that we stored the request `user_i$getFriendIDs()` in a variable called `user _i _friendsID` and we initialized a variable `total_ID`, before the loop, with an empty value. In the loop, for each $i = 1...10$ we stored the concatenation of `total_ID` and `user _i _friendsID` coming from each step $i = 1...10$ in the variable `total_ID`. The initialization before the loop is required if we want to avoid error message. For a better understanding, I illustrate the loop here after:



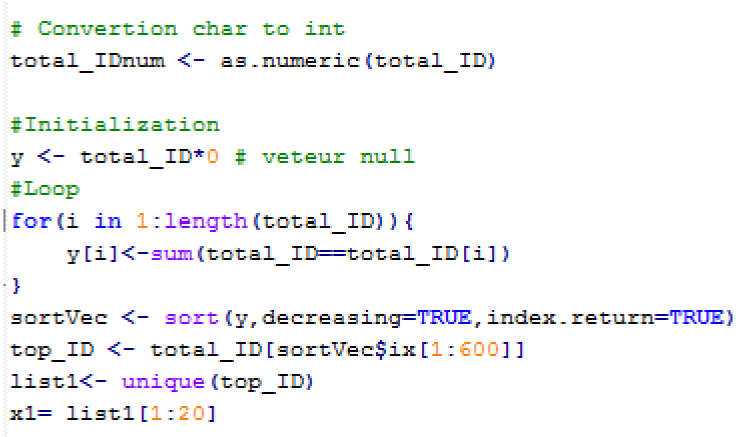
```
total_ID <- 0

for(i in 1:length(user_list)){
   #user_i : contain the user concerned by the iteration i
    user_i <- user_accounts[[i]]

    user_i_friendsID <- user_i$getFriendIDs() # friend IDs

    total_ID <- c(total_ID,user_i_friendsID)
}
```

Figure 3.2: Loop to extract data from a list of 10 Twitter accounts

Since Twitter API returns information extracted in a char type, we transformed all friend IDs in numerical value (integer in this case). At that point, we have a list of 6473 IDs stored in `total_ID` . To do this, we used the `as.numeric()` function with the variable `total_ID` as argument.

In order to obtain the 20 most common user IDs from the `total_ID` list, we assigned a score to each user. Each user's score is based on the fact that this user is followed by the other users of the `user_list`. For instance, if a user i is followed by 4 other users in the list, the score given to user i will be 4, knowing that all the friends of every user from the `user_list` were gathered in the `total_ID` list. The method can easily be programmed for a user i using the line `y[i] = sum(total_ID==total_ID[i])`. For each user i, we compare its ID to the database of the user's friend IDs. Each time the user i will be part of any other user's friend list, a score 1 will be added to the sum. Then, by sorting the score vector `y` in decreasing order, the top influencers of the `user_list` can be easily identified. This is done by using the `sort()` function with the three arguments: `y`, `decreasing=TRUE` and `index.return=TRUE`. The function returns two associated lists of components named `x` and `ix` containing the sorted numbers and the ordering index vector receptively. For example, $x[1] = 10$ and $ix[1] = 148$ rely on the ID with the index position 148 in the list `y` that appeared 10 times. We stored it in a variable called `sortVec`.

However, the list `y` still contains 6473 elements because we did not have unique elements in the list. For example, an ID that appeared 10 times in the `y` list would return 10 elements with the `sort()` function. To solve this problem, we first extracted the component ix and stored it in a new variable called `top_ID`. Then, we used the function `unique()`, with the variable `top_ID` as argument and stored the result a variable called `list1`. In order to obtain the 20 most common IDs, we selected the 20 first elements of `list1`. The code is illustrated as follows:



```
# Convertion char to int
total_IDnum <- as.numeric(total_ID)

#Initialization
y <- total_ID*0 # veteur null
#Loop
for(i in 1:length(total_ID)){
    y[i]<-sum(total_ID==total_ID[i])
}
sortVec <- sort(y,decreasing=TRUE,index.return=TRUE)
top_ID <- total_ID[sortVec$ix[1:600]]
list1<- unique(top_ID)
x1= list1[1:20]
```

Figure 3.3: R code to obtain the list of the 20 most influencer IDs

Iteration i

To reach 100 nodes, we have to repeat the process in 5 successive iterations. The structure of the code is the same as for the iteration 1 with few modifications. Note that in iteration i, we must extract all the friend IDs from the $list(i-1)$.

First, we have to adapt our code to the Twitter limitation of GET requests. Indeed, rate limits are divided into intervals of 15 minutes. This was a major issue during the implementation because when we tried to extract data with a loop, an error message appeared due to this limitation. We could solve this problem by dividing the list i in 2 and by using a `Sys.sleep()` function with 900 as argument (for 900 seconds). But, a better approach could be found by using the modulo operation. As we estimated that the system will block after 5 GET requests, we applied the following rule: for each i, if the i mod 5 is equal to 0, then `Sys.sleep()` function is to be used.

Secondly, when we obtained the variable `total_ID` of the iteration i, we concatenated it with the obtained `total_ID` in the previous iteration. For a better understanding, in iteration i, we qualified `total_ID` as the variable that stocks the previous concatenated list$(i-1)$ and `total_ID_it_i` as a variable containing the list of friend IDs obtained during the iteration i. When we obtained `total_ID_it_i`, we concatenated it with `total_ID` and stored it in a new `total_ID` variable. Then, we used `setdiff()` function instead of `unique()` because it sets the different IDs and returns a list with unique elements. The function `setdiff()` has two arguments : `top_ID` and its component x.

At the end of iteration 5, we obtained a list of 100 unique IDs which is the concatenation of list 1 to list 5 obtained during each iteration. We stored this list in a variable called `x`.

Construct the adjacency matrix 100x100

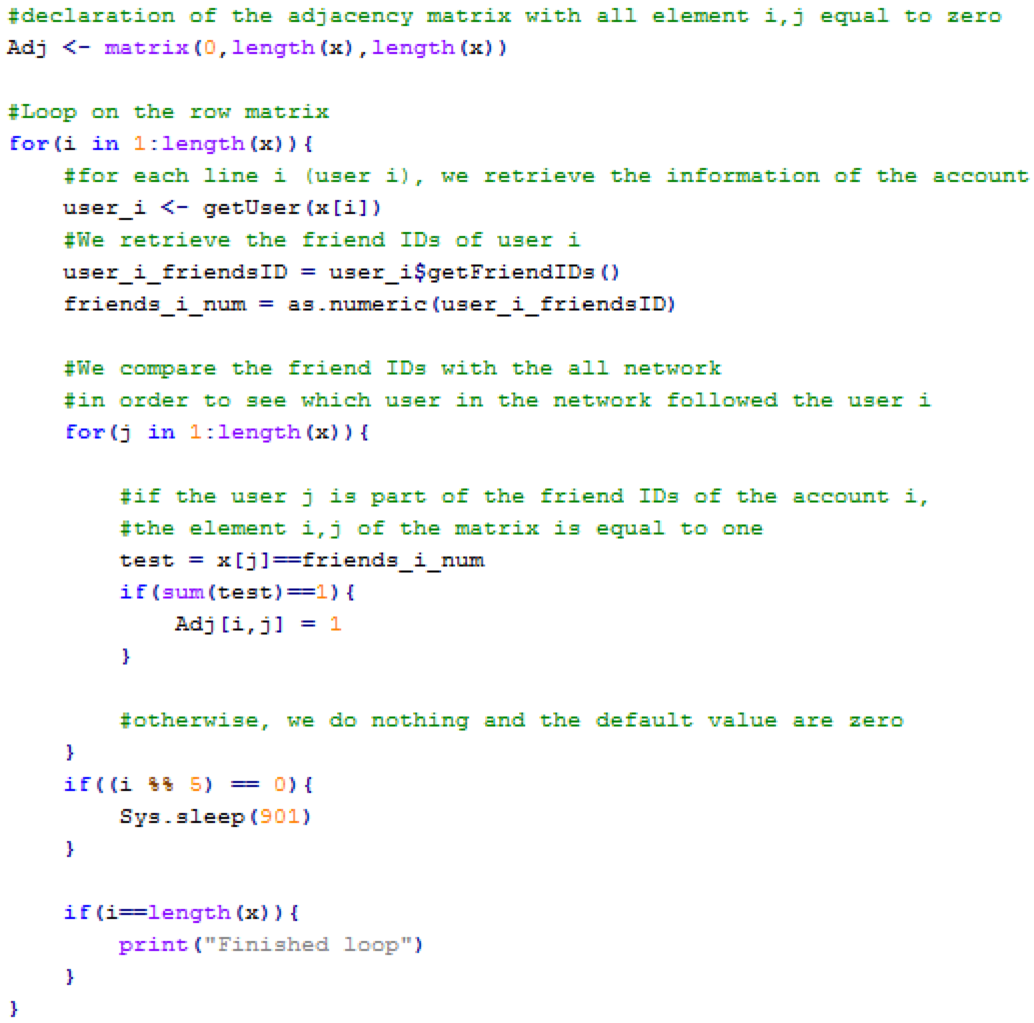
In order to make the relation between IDs of the 100 most common friends, we consider the adjacency matrix $A^{100\times100}$.

First, we have to declare the size of our adjacency matrix by using the function `matrix()` that contains three arguments : `0, length(x) and length(x)`. At the beginning, we initialized the matrix with all the entries $a_{ij} = 0$; this is why there is a parameter "0". The function `length()` determines the size of the list `x` (100 elements in this case). We stored all this information in a variable called `Adj`.

Now, consider X, the list of 100 elements $i = 1...100$ (stored in the variable `x`), as a vector column of the adjacency matrix A and Y the transpose of X with the same elements $j = 1...100$. For each element $i = 1...100$ in X, we retrieved their base information corresponding to the index i and stored it in a variable called `user_i`. In order to retrieve this information, we use the function `getUser()` with one argument: `x[i]` corresponding to the index i and stored it in a variable called `user_ID`. This function returns an object with the listed information of each user ID. Then, we extracted the friend IDs of the account i by using a GET request such as `user_i$getFriendIDs()`, we stored them in a variable named `user_i_friendsID` and, we transformed the type of this variable to numerical value.

When done, we compared all friend IDs of a user i with the all network in order to know which accounts in the network are followed by user i. Taking a user i, for each $j = 1...100$ in Y, if an element j is part of the friend IDs of user i, the entry a_{ij} will be equal to 1, otherwise, the entry a_{ij} will keep its original default value, i.e. 0.

Again here, we have to solve the Twitter API limitation of the GET request by the modulo technique described above. When the loop is finished, which means $i = 100$, the message "finished loop" is printed out. The code is described in the figure 3.4.



```
#declaration of the adjacency matrix with all element i,j equal to zero
Adj <- matrix(0,length(x),length(x))

#Loop on the row matrix
for(i in 1:length(x)){
    #for each line i (user i), we retrieve the information of the account
    user_i <- getUser(x[i])
    #We retrieve the friend IDs of user i
    user_i_friendsID = user_i$getFriendIDs()
    friends_i_num = as.numeric(user_i_friendsID)

    #We compare the friend IDs with the all network
    #in order to see which user in the network followed the user i
    for(j in 1:length(x)){

        #if the user j is part of the friend IDs of the account i,
        #the element i,j of the matrix is equal to one
        test = x[j]==friends_i_num
        if(sum(test)==1){
            Adj[i,j] = 1
        }

        #otherwise, we do nothing and the default value are zero
    }
    if((i %% 5) == 0){
        Sys.sleep(901)
    }

    if(i==length(x)){
        print("Finished loop")
    }
}
```

Figure 3.4: R code to obtain an adjacency matrix

3.3.3 Network Modelling with igraph in R

The package igraph (Csardi et al., 2015) is a collection of analysis tools for network data and it is open source and free. First, we have to install the package by the command `install.package ("igraph")` and then to load the package by the command `library(igraph)`.

First, we can plot the network obtained by creating the graph from our adjacency matrix. Stored in the variable `graph`, we can do it easily thanks to the `graph_from_adjacency _matrix()` function. This function contains three main arguments: `Adj` that stored our adjacency matrix; `mode` that specifies how igraph would interpret the supplied matrix (for a directed network, "mode" must be used as following : `mode="directed"`); and the last argument is used to specify if we want to create a weighted graph from our adjacency matrix or not (here, we have a binary relation and we write the argument as following: `weighted=FALSE`). This function will preserve the order of the vertex (the vertex corresponding to the first row will be the vertex 1 in the graph). This function also returns a graph object. By using the function `plot(graph)`, a directed and unweighted graph will appear (see figure 3.5).

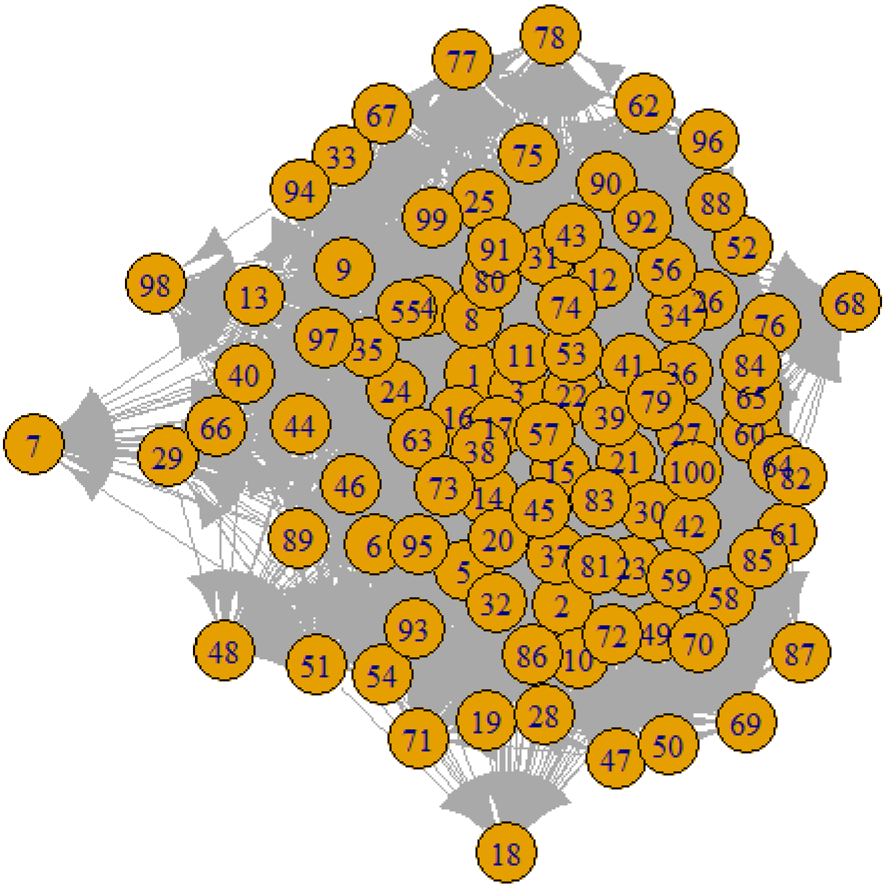


Figure 3.5: Graph translation of our adjacency matrix

Igraph package provides functions to analyse graphs but also functions that directly compute the centrality of a vertex.

Degree Centrality

The function `degree()` is available to know the DC of a vertex and takes at least four arguments. The first argument relies on the variable that stores our graph object to analyse. The second argument `mode`, can value `in` for indregree, `out` for outdegree. In this paper, as we computed the value for both indegree and outdegree centralities, the argument was used as following: `mode=c("in", "out")`. The third argument `loops=TRUE` means that the loop edges are also counted. The last argument is used to normalize the formulation: `normalized=TRUE`. In this way, we will obtain a numeric vector of degree scores for each nodes i.

Closeness Centrality

The function `closeness()` also needs four arguments closely similar to the previous one. The first argument is also the variable `graph`. As second argument we used `mode="in"` because we want to consider the path to a vertex rather than the path from a vertex (represented by "out"). The last ones are `weights=NULL` and `normalized=TRUE`. Normalization is performed by multiplying the raw closeness by $n - 1$, where n is the number of vertices in the graph. The function will return a numeric vector with the closeness values of all the vertices.

Betweenness Centrality

Again, the `betwenness()` function needs fours arguments: `graph`, `directed=TRUE`, `weights=NULL` and `normalized=TRUE`. It will return a numeric vector with the betweenness score for each vertex i.

Eigenvector Centrality

The function `eigen_centrality()` needs at least four arguments : `graph`, `directed=TRUE`, `weights=NULL` and `options=arpack_defaults`. This function will return a numerical vector of eigenvalue for each vertex. Note that `arpack_defaults` is an interface to the `arpack` library for solving large scale eigenvector problems.

Chapter 4

Results and Discussions

4.1 Graph Data Structure

After running the program 7 hours, an unweighted and directed graph with 100 nodes and 4881 edges was obtained. The `str()` function was used to print the `graph` information in a way that is easily readable.

The network is depicted in figure 3.5. Visually, the graph seems to be strongly connected, especially in the center of the graph. In fact, this high degree of connection reduced the distances between nodes: the center of the network is composed of nodes with a high indegree score and the extremity by nodes with a low indegree score. Also, the structure of the graph seems to be cyclic and there is no clear presence of any sub-group.

Moreover, the extraction process is based on the assumption that "friends from an account with large influence in a specific topic should also be an influencer and should talk about the same topic". Intuitively, to be part of the final sample, a member has to be considered as influencer in the fashion industry by the other members. This assumption avoids taking into account members with a high degree of passivity (Romero et al., 2011). Applying centrality measures on this sample provided a ranking of influencers between a sample of influencers. Obviously, all of these findings impact the results and I will try to signal when this impact occurs.

On the one hand, the characteristics of the sample are discussed and analysed to prove homophily. In the context of Twitter, homophily implies that a user follows another one because he is interested by its content and its friends follow him back because they share the same topics of interest.

First, based on the descriptions given by Twitter accounts, 53% of the sample referenced at least one of the following words in description: "fashion", "beauty", "wear" or "style". However, magazines and fashion houses generally do not mention these kind of words in description. Only 7 accounts have an empty description. In order to know exactly the users' topics of interest, each page was visited individually. Indeed, the sample is composed of a set of bloggers, designers, magazines, brands, or fashion e-commerce websites. I also distinguished that 89% of the accounts are interested in fashion topics. For the rest, it was difficult to know exactly their impact in fashion such as "KanyeWest" or "MattiasSwenson", the co-founder of Bloglovin.

Secondly, the reciprocity score of the network is equal to 0.62, which means that on average more than 62 % of the nodes are mutually linked. This measure is interesting because mutual links facilitate the transportation of information in a network. Due to the assumption made for the extraction of nodes, it is not surprising to find a high value for this parameter because common friends of a particular user in each iteration are taken into account. The clustering coefficient is an irrelevant measure here because only undirected graphs are considered.

Moreover, one should compare the number of followers and friends from our sample with their real number of followers and friends on the all Twitter. In the figure 4.1, one can see that most of the users do not follow back their followers because they are not especially interested in the content published by them. However, regarding the figure 4.2, this finding is not true in our sample where reciprocity exists. Our study confirms a previous study (Weng et al., 2010) stating that homophily exists in friendship relations and proves that users are serious when they are choosing friends. This is an important result in this study because it allows the identification of influencers using centrality measures. Otherwise, if users were not serious in following others, the sample in fashion topics would not be valid.

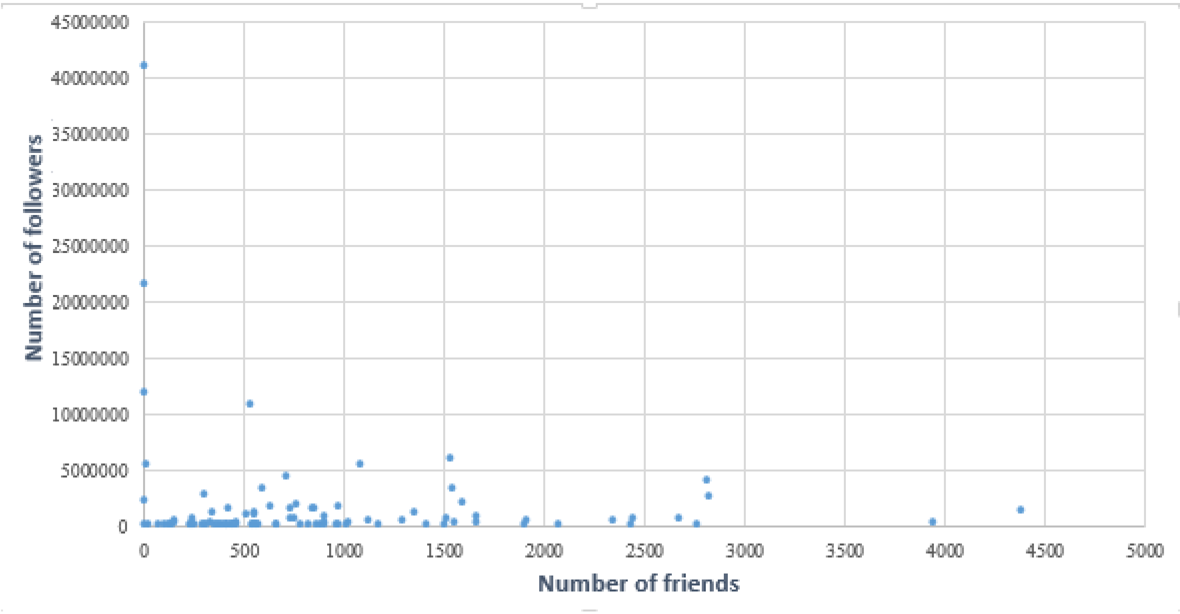


Figure 4.1: Number of friends vs number of followers on Twitter

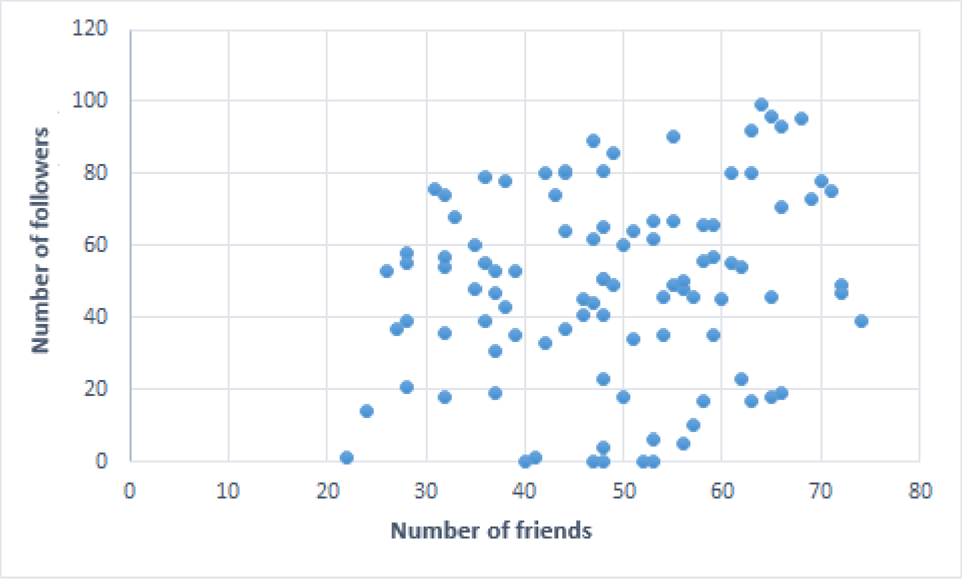


Figure 4.2: Number of friends vs number of followers in the sample

4.2 Centrality measures

In this study, the obtained results are ranked lists with the centrality scores of each Twitter account.

On the one hand, the first observation is that the ranking provided by Fashionista website and the ranking provided by centrality measures are not similar. This difference can be explained by the composition of the sample: different actors in the fashion industry are considered. Moreover, the survey was carried out in 2015, while Twitter provides real-time data. The Fashionsita survey is also based on other characteristics and canals such as Instagram or YouTube, while in this thesis, only network patterns on Twitter are considered.

On the other hand, correlation between rankings are depicted in table 4.1.

	Indegree	Outdegree	Closeness	Betweenness	Eigenvector
Indegree	1				
Outdegree	0.19	1			
Closeness	0.16	0.89	1		
Betweenness	0.5	0.75	0.68	1	
Eigenvector	0.98	0.15	0.12	0.43	1

Table 4.1: Correlations study on centrality scores

The results show varied correlations between rankings conform to the previous study discussed in chapter 2 (Valente et al., 2008).

First, indegree and outdegree results are not very correlated which means that a user with the highest number of friends is not necessarily the one with the highest number of followers and vice versa.

Indegree and eigenvector scores almost provide an identical ranking. Indeed, it is due to the computation of EC and how a node receives the influence from its inward neighbours. However, many studies on Twitter show that indegree measures the popularity or the size of the audience but not the degree of influence (Cha et al., 2010). For example, a user may buy followers to increase its popularity but it does not mean it is more influent. Moreover, this high correlation score also shows that the global structure of the network is not well considered in the EC results because all nodes are considered as influencers. In this specific situation, these two measures seem to be redundant but in the reality these measures present a lower correlation.

Eigenvector ranking is not very correlated with the outdegree one because the EC computation considers only incoming links.

Moreover, eigenvector, closeness and betweenness provide three different rankings. The non-redundancy of these measures gives the opportunity to discuss the results separately.

Betweenness and closeness are highly correlated with outdegree because both of these measures use a BFS algorithm to find the shortest paths. Also, users are highly connected with other members. For instance, 50 nodes out of 100 have a directed link to more than 55% of the other nodes.

Secondly, because indegree and outdegree centralities are highly correlated with some measures, we decided not to interpret their results in a fashion context. However, Appendices VIa and VIb depict the top10 ranking for indegree and outdegree centrality scores.

Closeness Centrality (CC)

The 10 accounts with the most influence based on the CC is shown in the Appendix V.

In the sample, the average centrality score is equal to 0.77 with a standard deviation equal to 0.09. There is no extreme value with a maximum at 1 and a minimum at 0.56. We can conclude that the majority of centrality scores is closed to the mean. Half of the nodes in the top10 derives from the first iteration during the extraction of nodes. However, this finding is not inversely true, the end of the ranking contains nodes from iterations 1 to 5.

Moreover, the presence of bloggers and of magazines are observed in the top10. Critics coming from these users are relevant for businesses in term of worth-of-month communication and it is very important for the fashion houses to be referenced by them. Nodes with the lowest centrality are mainly bands such as Louis Vuitton, Chanel or Kanye West.

Conceptually, the use of CC in our data shows a relevant measure to detect central nodes spending the minimum time to communicate with all others when they spread a message, according to Sabidussi (1966) point of view. Also in this study, with its huge number of links pointed to him, a user with a high CC score can spread rapidly information to others in the network. For instance, "thecoveteur" is connected with the complete network and 64% of nodes follow him back. Because this user is part of the direct predecessors list of the compete network, the CC score is equal to one.

Obviously, to reduce the complexity we can use outdegree measure rather that CC because of its correlation score. However in a context of Twitter, we can suppose that a member can easily increase his number of friends and therefore have a higher centrality score, but in reality it is not true. If someone decided to create a new account on Twitter and then to follow all of the 100 nodes, this user would not be considered in the extraction process because he is not part of the list of the other members previously extracted. Even if the outdegree and closeness are correlated in this case, we cannot replace one measure by the other one in the complete Twittosphere.

Betweenness centrality

The Appendix VI provides the top10 for this measure.

All of the betweenness scores are close to 0, which comes from the fact that the network structure and the properties of our sample are particular. Due to the high reciprocity scores, there is no pair-dependency on a node in the shortest paths. In other words, there is no boundary-spanner and the network has a cyclic structure. This can be justified by the assumption made when we extracted our nodes: a "friends of my friends are also my friends" strategy was considered and returned a network of 100 already influential persons in fashion with members highly connected together.

Then, one can also understand it based on the conception of the BFS. When the shortest path starting from a node s to a node t is computed, the number of predecessors obtained in the shortest path was low. This leads to a low pair-dependency value and in consequence a BC close to 0. I concluded that BC is not a relevant measure to quantify the importance of a node in our network. However, this interconnection between users should be reduced if the size of the network increase.

Eigenvector Centrality

The appendix VII provides the top10 ranking for this measure. In the results, all of eigenvalues are different in our ranking, which means our adjacency matrix is diagonalisable and valids the use of the power method. The EC mean is equal to 0.64 with a standard deviation equal to 0.18. The maximum value is 1 and the minimum value is 0.26. The top10 provided nodes from iterations 1 to 4 but the majority came from iterations 1 and 2. However, at the end of the ranking, half of nodes derived from iterations 4 and 5.

EC score is high if it has a highest number of links to other individuals. One can suggest that EC is a good indicator of activity. This measure is suitable to find well integrated users that are highly followed by many other influencers. It is not surprising to find magazines in the top10, because they are well integrated in fashion industry with strong connections with other influence nodes such as bloggers and brands. They gained credibility thanks to the fact that they highly reference other important fashion personalities or houses and in consequence are followed by them.

Similarly to closeness discussion, indegree and eigenvector must be considered separately in real world. In the sample, being followed by other influencers assumes that a user is also important by publishing interesting contents. However, indegree is not a good measure in the complete Twittosphere as discussed before.

4.3 Limits

First, remember at iteration 1, a list of 10 influencers were taken arbitrary and manually on Twitter in fashion industry. I decided not to consider their IDs in the first step because if these users were influent, then their IDs would appear in the next iterations. In our results, only three accounts (wendynguyen, Kayture, GalMeetsGlam) do not appear in our 100 nodes. Indeed, these three accounts do not figure in the friends list of our 100 nodes. Different hypotheses can be formulated at this step. First, the initial list of 10 bloggers also gained credibility through other social network sites and/or on their own website. For example, Wendy Nguyen is better known for her YouTube channel with videos that have been viewed 30 millions times. Secondly, these three persons may cover a segment in fashion or brands that are not covered by our 100 nodes. When our sample is extended to 200 nodes, these three users appeared in our final result.

If we come back to the definition of influencer in a marketing context, only influence based on network attributes is considered but other variables should also be integrated in our model.

Obviously, the final sample depends on our initial list of 10 influences persons. The robustness of this method should be tested to know if similarity exists for the final list when the initial list is modified. In an unreported test, I found a similarity of 82% to 91%. However, this result is not surprising because the fashion social networks are highly connected with high reciprocity between fashion influencers. Magazines and bloggers are already well integrated into their networks, compared to fashion houses which have a smaller number of friends. In order to test the performance of the mapping method, tests on other topics such as finance or politics must be done. Here, these two subjects are more similar and an accounts such as Barak Obama should probably appear in both topics because of his popularity. However, this limit can be solved by adding a manual selection between each extraction step. However, the time execution of the model will be higher.

Comparison should also be analysed between different social network sites. Some influencers may have more presence and impact on YouTube or Instagram than Twitter.

Secondly, our data structure is based on a followers/following network. As in the Bigonha's point of views (2010), this kind of configuration does not ensure that, when an influencer published a tweet or a picture, the others in the network received this information and are interested by its content. An interaction network constructed on retweets or mentions should be more suitable for that purpose. Also, precision in our topics-detection (e.g. fashion bloggers in France) should be feasible in interaction networks because of text, while it is not possible in followers/following networks.

Thirdly, our study provided to marketers a targeting strategy to optimize the diffusion of information by systematically targeting some classes of individuals but the relative cost of targeting the persons is still an unresolved empirical question. However, Bakshy, Homan & Mason (2011) provided a study on the relative cost of identifying and compensating potential influencers.

Finally, the choice of the centrality measures has also limits. As indicated, the communication is not only diffused in the geodesic paths and EC considers the external sources of information to quantify the influence of an actor. Other measures cited should be more adapted to real situations.

Chapter 5

Conclusion

Communicative behaviour has changed through the growth in web-based social networks. This induces an increasing number of companies which are interested in the use of such networks to market a product or a service. The identification of actors who are well integrated into the network became a major issue. For this purpose, many centrality measures have been developed and analysed in recent years in this domain. Due to the increasing importance of social network sites such as Twitter, this paper aimed to:

(1) present the current state of research on degree, closeness, betweenness and eigenvector centrality measures in unweighted but directed and undirected social networks,

(2) provide to agents in charge of communication in a company a new and quicker way to extract Twitter data with a given topic and based on a friendship network,

(3) apply centrality measures on fashion topic which has not been developed in the literature.

Since there is not agreement on what centrality is, new measures are proposed everyday. Here, the provided state of research gives a better comprehension and proposes a better selection depending on what we want to measure: centrality as control (BC), centrality as independence (CC) or centrality as activity (DC and EC). In addition, this study compares the robustness and properties of the four basic concepts through several conceptual and empirical studies. For example, eigenvector centrality shows great robustness and accuracy because it takes the overall structure of the network in its calculation, while betweenness centrality is the less robust but is really performant to detect boundary-spanners in a network. Moreover, depending on the application cases, each measure must be interpreted regarding the context and the data sampled.

In a second part, I crawled 100 nodes and I obtained 4881 directed relations between actors in the fashion Twittersphere. The major finding is the presence of homophily in our sample that valids our extraction method for applying centrality measures. In our sample, users chose their friends in a serious way because they are interested by the content they are publishing. However, the method shows limits when we want to be more precise in the topic-detection. Interaction network seems to be more adapted for this case (see Moro et al., 2010).

The last contribution was applying the four basic concepts discussed in fashion, still undeveloped in the literature. However, the extraction method already provides a sample of influencers and the centrality measures allow to rank them.

In conclusion, for someone in charge of the communication in a fashion house, I advise in-closeness centrality as a good measure to find users that spend the minimum time to communicate with all others when they spread a message, while the BC is not adapted to the data used in this thesis because there are any presence of sub-group in the graph. EC shows great characteristics to find magazines or accounts highly referenced and connected with other important nodes. Other empirical studies should be tested with centrality measures closer to the real world in order to compare the obtained rankings.

Finally, this thesis is the first step in the analysis of online social networks in HEC Liege. Plenty of applications may be created to make use of the huge information available on the Internet. Obviously, the model developed here is not perfect but we can be proud of providing a new and easily way to map a network based on friendship connections with a given topic. Future research should be carried out on interaction networks in order to compare it with our results.

.1 Appendix I: Approaches to the analysis of centrality measures (CM)

Authors	Approach	Analysed CM
Freeman et al., (1979)	Interpretation of different concepts of centrality and application on it	DC, BC and CC
Valente et al., (2008)	Analysis of the stability of CM in sampled network	DC, CC, BC and EC
Bolland (1988)	Analysis of performance of the four basic concepts through an empirical study in real and simulated networks	DC, BC, CC and EC
Borgatti (2006)	CM robustness in case of imperfect data	DC, BC, CC and EC
Costenbader & Valente (2003)	Analysis of the robustness when the networks are sampled	DC, BC, CC and EC
Friedl & Heidemann (2010)	Assessment of properties of CM in a conceptual study	DC, BC, CC and EC

Authors	Approach	Analysed CM
Hossain et al., (2007)	Diffusion of message through influencer in telecommunication networks	DC, BC, CC and EC
Lee et al., (2010)	Analysis of CM in a customer database	DC and BC
Yan & Ding (2009); Mutschke (2003)	CM in co-authorship network	CC, BC and DC
Bichler (2008)	CM to detect influencers who achieve a maximum dissemination of a marketing message in Telecom	outdegree
Cha et al., 2010	Indegree, retweets and mentions influence on Twitter	DC
Romero et al., 2011	Influence and passivity on Twitter	DC and EC
Bigonha et al., 2010	Detection of evangelists and detractors on the brevery sector in Brazil on Twitter	DC, BC and EC
Weng et al., 2013	Method for detecting topic-sensitive influential twitterers in friendship networks	DC and EC
Freeman et al., (1991)	Flow betwenness	related to BC
Newman (2003)	K-path centrality	related to BC
Tran (2014)	Hierarchical closeness	related to CC
Bonacich et al., 2001	Alpha centrality	related to EC
Page et al., 1999	PageRank	related to EC
Katz (1953)	Katz’s centrality	related to EC

Table 1: Appendix I: Approaches to the analysis of centrality measures (CM)

.2 Appendix II: List of 10 influence bloggers in Fashion from Fashionista magazine

Index	Name	User name	Descriptions
1	Chiara Ferragni	@ChiaraFerragni	She is 27 year-old and lives in Los Angeles(L.A). She is a global star and has more than 300 000 followers on Twitter.
2	Aimee Song	@AIMEESONG	She is 28 and lives also in L.A. She is an interior designer whose gained her popularity through street-style shots. She has more than 70 000 followers on Twitter.
3	Wendy Nguyen	@wendynguyen	Again in L.A, she is also famous on YouTube with some videos viewed over than 29 million times.
4	Kristina Bazan	@Kayture	She is a Swiss top model and with only 21 year-old. She is working with famous brands such as Louis Vuitton, Hugo Boss and Piaget.
5	Zanita Whittington	@zanitazanita	She is 28 and lives in Stockholm. She gained credibility through modelling, bolling and photography advice.

Index	Name	User name	Descriptions
6	Rumy Neely	@rumineely	She is 31 and has collaborated with established brands. However, she launched her own line of slip dresses, tap pants and distinctly cut tees.
7	Nicole Warne	@garypeppergirl	She is 25 and already business woman. She shows clothes in order to market them in her e-commerce website. She is one of the largest online vintage retailers in Australia.
8	Blair Eadie	@BlairEadieBEE	She is 29 and uses her blog to act as a merchandiser for brands like Gap and Tory Burch
9	Julia Engel	@GalMeetsGlam	She is 23 and lives in San Francisco. She is also well known in the fashion industry for shop affiliate links.
10	Nicolette Mason	@nicolettemason	This 29 year-old blogger is known for her advice on collections and discussion on bigger issues. She also writes a monthly column for *Marie Claire* magazine and has built also a TV personality.

Table 2: Appendix II: List of 10 influence bloggers in Fashion from Fashionista magazine

.3 Appendix III: List 1 of 20 most common friends

Index	ID	User name
1	228379737	"The Coveteur"
2	136361303	"Vogue Magazine"
3	19212009	"Who What Wear"
4	64822927	"Jane Aldridge"
5	73359920	"Valentino"
6	21190774	"Jessica Stein"
7	17809182	"Mattias Bloglovin’"
8	16271952	"Rebecca Minkoff"
9	54819492	"Liz Cherkasova"
10	437086963	"3.1 Phillip Lim"
11	21208444	"Barneys New York"
12	141598384	"Man Repeller"
13	19390810	"Sincerely Jules"
14	267947445	"NET-A-PORTER"
15	15934926	"KCD"
16	24190981	"Lucky Magazine"
17	44084633	"Teen Vogue"
18	46470446	"Louis Vuitton"
19	250314584	"Tommy Ton"
20	7092102	"bryanboy"

Table 3: Appendix III: List 1 of 20 most common friends

.4 Appendix IVa: Top 10 of the highest outdegree score of our 100 nodes

Node	Username	Outdegree score	Twitter description
1	"thecoveteur"	0,99	"Behind-the-scenes & beyond."
38	"susiebubble"	0,96	"A girl who likes a good chunky heel but hates chunky cheese."
3	"WhoWhatWear"	0,95	"Fashion and style, decoded."
17	"TeenVogue"	0,93	"Future tastemakers start here. Snapchat: teenvogue"
11	"BarneysNY"	0,92	"As Sarah Jessica Parker once told Vanity Fair, 'If you're a nice person and you work hard, you get to go shopping at Barneys. It's the decadent reward.'"
16	"LuckyMagazine"	0,90	"Read it. Love it. Shop it."
73	"atprettybirds"	0,89	"Blogger, Photographer, Contributing Style Editor at Lucky Magazine, Director of Style and Digital Content at Out There Creative Agency."
57	"saks"	0,86	"Tweeting from inside the most style obsessed place in the world. For service questions, tweet @saksservice or call 1-877-551-7257."
15	"KCDworldwide"	0,81	"The latest news from inside KCD, the leading fashion public relations, production and digital agency worldwide."

Table 4: Appendix IVa: Top 10 of the highest outdegree score of our 100 nodes

.5 Appendix IVb: Top 10 of the highest indegree score of our 100 nodes

Node	Username	Indegree score	Twitter description
2	"voguemagazine"	0,74	"The official twitter page of Vogue Magazine,"
23	"VogueRunway"	0,72	"See fashion first,"
12	"ManRepeller"	0,72	"Where an interest in fashion never minimizes one's intellect,"
22	"Refinery29"	0,71	"Read it. Love it. Shop it."
41	"TheCut"	0,70	"Fashions, Fame, Beauty, Goods, Love & War, Life is a runway,"
39	"Fashionista_com"	0,69	"All the fashion news you need to know,,, fast,"
3	"WhoWhatWear"	0,68	"Fashion and style, decoded."
60	"wwd"	0,66	"Fashion, Beauty, Business,"
57	"IntoTheGloss"	0,66	"Into The Gloss is a website dedicated to beauty,"
17	"TeenVogue"	0,66	"Future tastemakers start here. Snapchat: teenvogue"

Table 5: Appendix IVb: Top 10 of the highest indegree score of our 100 nodes

.6 Appendix V: Top 10 of the highest closeness score of our 100 nodes

Node	Username	Closeness score	Twitter description
1	"thecoveteur"	1	"Behind-the-scenes & beyond."
38	"susiebubble"	0,9705882	"A girl who likes a good chunky heel but hates chunky cheese."
3	"WhoWhatWear"	0,961165	"Fashion and style, decoded."
17	"TeenVogue"	0,9428571	"Future tastemakers start here. Snapchat: teenvogue"
11	"BarneysNY"	0,9339623	"As Sarah Jessica Parker once told Vanity Fair, 'If you're a nice person and you work hard, you get to go shopping at Barneys. It's the decadent reward.'"
16	"LuckyMagazine"	0,9166667	"Read it. Love it. Shop it."
73	"atprettybirds"	0,9082569	"Blogger, Photographer, Contributing Style Editor at Lucky Magazine, Director of Style and Digital Content at Out There Creative Agency."
57	"saks"	0,8839286	"Tweeting from inside the most style obsessed place in the world. For service questions, tweet @saksservice or call 1-877-551-7257."
15	"KCDworldwide"	0,8461538	"The latest news from inside KCD, the leading fashion public relations, production and digital agency worldwide."

Table 6: Appendix V: Top 10 of the highest closeness score of our 100 nodes

.7 Appendix VI: Top 10 of the highest betweenness score of our 100 nodes

NODE	Username	Betweenness	Description
38	"susiebubble"	0,016434403	"A girl who likes a good chunky heel but hates chunky cheese."
1	"thecoveteur"	0,0162506325	"Behind-the-scenes & beyond."
59	"BoF"	0,01478392	"Follow for breaking news and fashion business intelligence. Home of the #BoF500. Subscribe to our daily newsletter: http://t.co/67Jc1TCEbw"
17	"TeenVogue"	0,014482768	"Future tastemakers start here. Snapchat: teenvogue"
73	"atprettybirds"	0,014467044	"Blogger, Photographer, Contributing Style Editor at Lucky Magazine, Director of Style and Digital Content at Out There Creative Agency."
3	"WhoWhatWear"	0,013493541	"Fashion and style, decoded."
11	"BarneysNY"	0,013358209	"As Sarah Jessica Parker once told Vanity Fair, 'If you're a nice person and you work hard, you get to go shopping at Barneys. It's the decadent reward.'"
20	"bryanboy"	0,011561009	"Fashion blogger"
41	"TheCut"	0,011500643	"Fashions, Fame, Beauty, Goods, Love & War. Life is a runway."

Table 7: Appendix VI: Top 10 of the highest betweenness score of our 100 nodes

.8 Appendix VII: Top 10 of the highest eigenvector score of our 100 nodes

NODE	Username	Eigenvector	Description
2	"voguemagazine"	1	"The official Twitter page of Vogue Magazine."
23	"VogueRunway"	0,9783213	"See fashion first."
41	"TheCut"	0,9517799	"Fashions, Fame, Beauty, Goods, Love & War. Life is a runway."
12	"ManRepeller"	0,9473364	"Where an interest in fashion never minimizes one's intellect."
39	"Fashionista_com"	0,9219002	"All the fashion news you need to know... fast."
22	"Refinery29"	0,913945	"R29 unfiltered, uncensored — the best in fashion, health, entertainment, beauty, news... oh, and GIFs. https://t.co/s9eo6iDDGD"
60	"wwd"	0,8940097	"Fashion. Beauty. Business."
21	"IntoTheGloss"	0,8912001	"Into The Gloss is a website dedicated to beauty."
3	"WhoWhatWear"	0,8780607	"Fashion and style, decoded."
64	"TheLSD"	0,8763018	"Vogue Magazine, Contributing Editor / Moda Operandi, Founder"

Table 8: Appendix VII: Top 10 of the highest eigenvector score of our 100 nodes

Appendix A

References

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America, 101(11)*, 3747-3752.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *Journal of the acoustical society of America.*

Bavelas, A. (1948). A mathematical model for group structures. *Human Organization 7*, 16-30.

Bass, F. M. (2004). Comments on "a new product growth for model consumer durables the bass model". *Management science*, 50 (12_supplement), 1833-1840.

Beauchamp, M. A. (1965). An improved index of centrality. *Behavioral Science 10*, 161-163.

Benzi, M., Klymko, C. (2014) A matrix analysis of different centrality measures.*arXiv preprint arXiv:1312.6722.*

Bigonha, C., Cardoso, T. N., Moro, M. M., Almeida, V. A., & Goncalves, M. A. (2010). Detecting evangelists and detractors on Twitter. *In 18th Brazilian Symposium* *on Multimedia and the Web*, 107-114.

Bolland JM (1988). Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social Networks 10(3)*, 233-253

Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, 1170-1182.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social networks*, 29(4), 555-564.

Booth, N., & Matic, J. A. (2011). Mapping and leveraging influencers in social media to shape corporate brand perceptions.*Corporate Communications: An International Journal*, 16(3), 184-191.

Borgatti, S. P. (2006). Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, 12(1), 21-34.

Borgatti, S. P., Jones, C., & Everett, M. G. (1998). Network measures of social capital. *Connections*, 21(2), 27-36.

Borgatti SP, Carley KM, Krackhardt D (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks* 28(2), 124-136

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *science*, 323(5916), 892-895.

Brandes, U. (2001). A faster algorithm for betweenness centrality*. *Journal of mathematical sociology*, 25(2), 163-177.

Burt, R. S. (1997). The contingent value of social capital. *Administrative science quarterly*, 339-365.

Ronald, B. (1992). Structural holes: The social structure of competition. *Cambridge: Harvard.*

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems, 1695(5)*, 1-9.

Castells, M. (2004). Informationalism, networks and the network society: a theoretical blueprint [on-line resource]. *Castells M. The Network Society: a cross-cultural perspective/M. Castells.–Northampton, MA: Edward Elgar.*

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM*, 10(10-17), 30.

Costenbader, E., & Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social networks*, 25(4), 283-307.

Cross, R., Borgatti, S. P. & Parker, A. (2001). Beyond answers: dimensions of the advice network. *Social networks*, 23(3), 215-235.

Dalgaard, P. (2008). Introductory statistics with R. *Springer Science & Business Media.*

Dijkstra, E.W., 1959. A note on two problems in connexion with graphs. *Numerische Mathematik 1*, 269-271.

Easley, D., & Kleinberg, J. (2010). Networks, crowds, and markets: Reasoning about a highly connected world. *Cambridge University Press.*

Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), p. 210-230

Feng, P. E. B. J. (2011). Measuring user influence on Twitter using modified k-shell decomposition. *In ICWSM '11 Workshops*, 18-23

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.

Freeman, L. C., Roeder, D., & Mulholland, R. R. (1979). Centrality in social networks: II. Experimental results. *Social networks*, 2(2), 119-141.

Freeman L. C.(1979). Centrality in Social Networks: Conceptual Clarification. *Social Networks*, 215-239

Friedl, D. M. B., & Heidemann, J. (2010). A critical review of centrality measures in social networks. *Business & Information Systems Engineering* ,2(6), 371-385.

Gayo-Avello, D. (2013). Nepotistic relationships in Twitter and their impact on rank prestige algorithms. *Information Processing & Management*, 49(6), 1250-1280.

Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F.,& Gummadi, K. P. (2012). Understanding and combating link farming in the Twitter social network. *In Proceedings of the 21st international conference on World Wide Web*, 61-70.

Gentry J. (2011). Package 'twitteR'. Available on line in
http://cran.r-project.org/web/packages/twitteR/twitteR.pdf.

Hossain, L., Chung, K. S. K., & Murshed, S. T. H. (2007). Exploring temporal communication through social networks. *In Human-Computer Interaction-INTERACT 2007.* Springer Berlin Heidelberg, 19-30

Ibarra, H., & Andrews, S. B. (1993). Power, social influence, and sense making: Effects of network centrality and proximity on employee perceptions. *Administrative science quarterly*, 277-303.

Katz, E., & Lazarsfeld, P. F. (1955). Personal Influence, The part played by people in the flow of mass communications. *Transaction Publishers.*

Katz, L., (1953). A new status index derived from sociometric analysis. *Psychometrika 18*, 39-43.

Keller, E., & Berry, J. (2003). *The influentials: One American in ten tells the other nine how to vote, where to eat, and what to buy.* Simon and Schuster.

Kiss, C., & Bichler, M. (2008). Identification of influencers-measuring influence in customer networks. *Decision Support Systems*, 46(1), 233-253.

Knoke, D., & Yang, S. (2008). *network analysis* (Vol. 154). Sage.

Knoke, D. (1982). The spread of municipal reform: Temporal, spatial, and social dynamics. *American lournal of Sociology. 87*, 1314- 1349.

Kozinets, R. V., De Valck, K., Wojnicki, A. C., & Wilner, S. J. (2010). Networked narratives: Understanding word-of-mouth marketing in online communities. *Journal of marketing*, 74(2), 71-89.

Kratzer, J., & Lettl, C. (2009). Distinctive roles of lead users and opinion leaders in the social networks of schoolchildren. Journal of Consumer Research, 36(4), 646-659.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media?. *In Proceedings of the 19th international conference on World wide web* , 591-600

Lee, S. H. M., Cotte, J., & Noseworthy, T. J. (2010). The role of network centrality in the flow of consumer influence. *Journal of Consumer Psychology*, 20(1), 66-77.

Leskovec J, Horvitz E (2008) Worldwide buzz: Planetary-scale views on a large instant-messaging network. *In Proceedings of the 17th international conference on World wide web* , 591-600

Mehra, A., Kilduff, M., & Brass, D. J. (1998). At the margins: A distinctiveness approach to the social identity and social networks of underrepresented groups. *Academy of Management Journal*, 41(4), 441-452.

Milgram, S. (1967). The small world problem. *Psychology today*, 2(1), 60-67.

Mutschke, P. (2003). Mining networks and central entities in digital libraries. A graph theoretic approach applied to co-author networks. *Advances In Intelligent Data Analysis V*, 2810, 155-166

Nieminen, J. (1974). On the centrality in a graph. *Scandinavian journal of psychology*, 15(1), 332-336.

Newman, M. E. (2004). Analysis of weighted networks. *Physical Review E*, 70(5), 56.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2), 167-256.

Newman, M. E. (2005). A measure of betweenness centrality based on random walks.*Social networks*, 27(1), 39-54.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3), 245-251.

Panju M. (2011) Iterative methods for computing eigenvalues and eigenvectors, *The Waterloo Mathematics Review*, 1(1): 9-18.

Parlett, B. N. (1974). The Rayleigh quotient iteration and some generalizations for nonnormal matrices. *Mathematics of Computation*, 28(127), 679-693.

Riquelme, F. (2015). Measuring user influence on Twitter: A survey. *arXiv preprint arXiv:1508.07951.*

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). Influence and passivity in social media. *In Machine learning and knowledge discovery in databases* pp. 18-33.

Scott J (2000). Social Network Analysis: A Handbook. *2nd. London:: Sage Publications.*

Scott J (2011) Social network analysis: developments, advances, and prospects. *In: Social network analysis and mining*, 1-6

Scott, J. (2012). Social network analysis. *Sage.*

Sernovitz, A., & Kawasaki, G. (2006). Word of mouth marketing. *USA: Dearborn Trade, A Kaplan Professional Company.*

Tran, T. D., & Kwon, Y. K. (2014). Hierarchical closeness efficiently predicts disease genes in a directed signaling network. Computational biology and chemistry, 53, 191-197.

Valente T (1996) Social network thresholds in the diffusion of innovations. *Social Networks*, 18(1), 69-89

Valente, T. W. (1995). Network models of the diffusion of innovations. *Computational & Mathematical Organization Theory*, 2(2), 163-164.

Valente, T. W., Coronges, K., Lakon, C., & Costenbader, E. (2008). How correlated are network centrality measures?. Connections (Toronto, Ont.), 28(1), 16.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications (Vol. 8).* Cambridge university press.

White, D. R., & Borgatti, S. P. (1994). Betweenness centrality measures for directed graphs. *Social Networks*, 16(4), 335-346.

Weimann, Gabriel (1991), The Influentials: Back to the Concept of Opinion Leaders?,*Public Opinion Quarterly*, 55, 267-79.

Wen Chai, Wei Xu, Meiyun Zuo, & Xiaowei Wen (2013) A novel framework to identify and predict influential users in micro-blogging. *In Jae-Nam Lee, Ji-Ye Mao, and James Y. L. Thong, editors, 17th Pacific Asia Conference on Information Systems, PACIS 2013, Jeju Island, Korea*, 18-20-22

Weng, J., Lim, E. P., Jiang, J., & He, Q. (2013). Twitterrank: finding topic-sensitive influential twitterers. *In Proceedings of the third ACM international conference on Web search and data mining*, 261-270

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, 60(10), 2107-2118.

A.1 Executive Summary

Online social networks have facilitated the interaction and topic discussion. Some of this content became a rich and important source of information and also strategical for companies. One of the most popular of such websites is Twitter.

Nowadays, the use of digital influencers became a new strategy in the development and in the management of marketing campaigns for leading brands and companies. Fashion industry usually targets them to market products or to diffuse messages. In consequence, the identification of these persons became a central issue for marketers.

In this dissertation, I propose a state of research in centrality measures, developed in social network analysis, in order to identify those influencers. Interpretability, robustness and accuracy, current applications and related work on Twitter will be discussed in order to select and understand these concepts. Moreover, I propose a new technique to collect Twitter data with a friendship graph and with a given topic. I perform this research on fashion industry which has not been treated yet in the literature, and then, I use centrality measures to identify the most influential users. The experimental evaluation shows that the presence of reciprocity can be explained by phenomenon of homophily. This finding valids the extraction process to create a sample composed of users interested and influent in fashion topics. The application of centrality measures on the sample provides a relevant ranking of influencers that can be used in a marketing campaign.

Keywords: Twitter, Centrality measures, Social network analysis, Degree centrality, Closeness centrality, Betweenness centrality, Eigenvector centrality, Influencer, Network typology, Digital influencer marketing