Introduction

Social networks has attracted the attention of researchers, companies, politicians, and governmental institution these recent years. This gain of interests is due to the fact of the increasing applications targeted by the aforementioned actors. Effectively, the huge amount of data exchanged through these networks is analyzed to the purpose of several applications such as propaganda, business intelligence, defense, etc. This growing phenomena becomes an important part of our everyday life, millions of people exchange a tremendous amount of data on Tweeter, Facebook, Instagram, etc. These free micro-blogging services, allow people to register, broadcast, share content and follow other members. Collaborative networks is another type of social networks where members collaborate to achieve common goals, examples of such sites includes Database List of Publications (DBLP), ResearchGate, etc.

The fast evolution of the social networks such as Twitter, Facebook represent an important factor in social influence emergence. Influencers are a complex force in the internet and are considered as propaganda agents in the virtual world. Discovery of influential members is important in various domains such as viral marketing and political campaigns.

Social influence has been a widely accepted phenomenon in social networks for many decades, in recent years researchers have developed a variety of techniques and models to capture information diffusion in online social networks, predict it, and extract knowledge data to discover spreaders in such domain. Two main methods for influencers detection were identified in the literature. The first type of methods employs the structure of the network for this purpose, whereas the second category is based on the content of interactions.

Another very active subject is the study of information propagation through this type of media.The propagation models in social networks (in which users share content) seek to replicate the dissemination of information between users. Being able to model this well diffusion allows their use in many applications, such as identification of influence nodes, or the identification of links to remove or limit the spread (for example to stop the rumors). We are interested in this manuscript in the detection of influencer nodes and more particularly to the real information epicenter. When one wants to disseminate content in a social network, people often try to reach as many people. To do this, the problem seeks the initial broadcasters that will result in maximum spread. To understand more what makes someone an influencer is when one's opinions, emotions, or behaviors are affected by others, intentionally or unintentionally. The forms of influence refers by an individual as evident by their participation in the

conversation and in global influence refers to how influence spreads among the social network. In this work we proposed a new approach based on a previous work on influencers detection using information cascade, that enables to discover influencers based on content of interaction and the underlying graph.

This manuscript is organized as follows, in the first chapter we introduce some general concepts and definition in order to introduce the important concepts of the social network analysis. In the second chapter we have presented a brief state of art of influencers detection, we have defined the notion of social influence, its context and why their detection is essential. Afterwards, we have focused on recent works in this domain. Finally, in the third chapter we have presented our approach which is based on a previous work that uses information cascade to detect most influencers people in networks.

I.1 Introduction

Social networks has attracted the attention of researchers, companies, politicians, and governmental institution these recent years. This gain of interests is due to the fact of the increasing applications targeted by the aforementioned actors. Effectively, the huge amount of data exchanged through these networks is analyzed to the purpose of several applications such as propaganda, business intelligence, defense, etc. This growing phenomena becomes an important part of our everyday life, millions of people exchange a tremendous amount of data on Tweeter, Facebook, Instagram, etc. These free micro-blogging services, allow people to register, broadcast, share content and follow other members. Collaborative networks is another type of social networks where members collaborate to achieve common goals, examples of such sites includes Database List of Publications (DBLP), ResearchGate, etc. All the aforementioned examples of social networks represent a map of specific ties between entities which exhibit an interdependency such as friendship, communication interests, financial exchange, etc.

To make this manuscript self-contained, we introduce in this chapter social networks and some of their core concepts.

I.2 What is a social network?

Basically, a social network is defined as a network of interactions or relationships where nodes represent actors and edges consist of relationships. Formally, a social network is considered as a graph $G = (V, E)$, where V represents a set of nodes also called vertices and E stands for the set of edges connecting pairs of nodes. The network structure answers the question who is related to whom. Networks could be explicit such as friend network, or could be the result of the activities of the members such as publishing in DBLP.

Social graphs could be directed, or undirected according to the relation between nodes. If this relation is symmetric the graph is undirected, otherwise it is a directed graph. For instance, an email network is a directed graph. Effectively, links exhibit a given direction since a relation between two members illustrates the fact that one of them had sent an email to the other one. An undirected graph is connected if some path connects every two nodes in the network, the components of a graph (or network) are the distinct maximally connected sub-graphs, while, a directed graph is connected if the underlying undirected graph is connected (i.e., ignoring the directions of edges). Graphs could also be weighted or unweighted, in the

first case the weights of ties represent the between-nodes link strength. Figure 1.1 below depicts different types of social graphs.

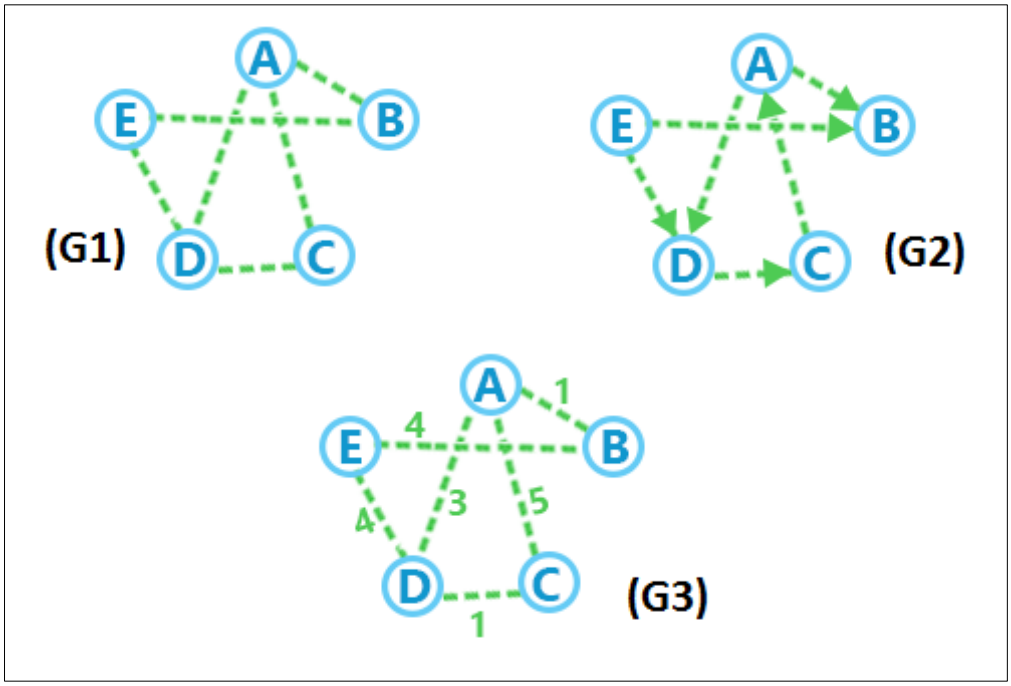


Figure 1.1 Illustration of social graphs types **(G1)** is an undirected and unweighted graph, **(G2)** is a directed graph, **(G3)** is a weighted graph.

I.2.1 Nodes and social graph properties

Nodes in social graphs are characterized by mean of several properties. Some metrics were designed to assess the importance of a node in a graph such as:

- **Betweenness centrality**: This property was first introduced in sociology to define social weight of a node [1]. It is a very important notion in graph theory, this property is based on shortest paths between nodes in the graph. The betweenness of a node N_i or an Edge E_i is the total number of shortest paths between all pairs of nodes in the network passing through the N_i or E_i.

- **Degree centrality***:* It is a measure of node popularity in a network, it is defined according to the number of edges connected to a particular vertex in the network, and is formally defined as follows:

$$D_c(i) = \frac{d_g(i)}{N-1} \quad (1)$$

where N stands for the total number of nodes in the network, $d_g(i)$ is the degree of node i. A node degree is simply the number of connections that a node has. On Twitter, it is the number of followers; on Facebook, it represents the number of friends.

- Neighborhood and degree of node: the neighborhood of a node i is the set of nodes that i is connected to, in an undirected graph the degree of node is the number of edge that involve it, whereas for the directed graph the degree is the sum of the indegree (number of incoming edges) and outdegree (number of outgoing edges).

In addition to nodes metrics, some measures were introduced to indicate characteristics of social graphs such as:

- Diameter: It is defined as the largest distance between any two nodes in the network. Given $D(i, j)$ the length of the shortest path between nodes i and j, the diameter represents the largest value of D for each node in network.

- Average path length: it stands for the average distance between any two nodes in the social network.

I.2.2.1 Twitter

In March 2006, a little communications service called ***Twttr*** debuted. It began as a side project at a San Francisco podcasting company, but it was not long before the side project became the main event. Today, just ten years later, Twitter is booming. The service has more than 310 Billion users, it supports over 35 languages, 500 million tweets are sent per day.Twitter is a free micro-blogging service, users can subscribe, share instantly content, and follow other members.

I.2.2.2 Twitter-specific terms

- **Time line:** is a screen that displays a stream of **Tweets** from followed Twitter members, user can **Reply**, **Retweet** or **Like** a Tweet from within the timeline.
- **Posts:** are any message published on Twitter which may contain photos, videos, links. Twitter users call their posts tweets. Unlike Facebook or LinkedIn tweets are searchable and public, they must be small to be sent as a text message to followers, a tweet must be less than 140 characters, members can add hashtags to a keyword in their post.
- **Hashtag:** is a term, prefixed by the **#** symbol, Hashtags are used to signify that a post is about a particular topic to allow for easy searching.

- **Follow:** when you are interested to other twitter users, you subscribe to their twitter update. With this subscribing action you will be a follower to them, consequently, every time they post a new tweet, it will show up on your Twitter time line.
- **Retweet:** is the act of posting a tweet from someone you follow and you share with your followers.
- **Direct message:** is a private message sent to you, or by you, over Twitter.

I.3 Properties of social networks

In what follows, we introduce some essential properties of social networks that have emerged from studies and experimental works in several disciplines.

I.3.1 Small world phenomenon

What is commonly known as small-world effect [2], originates from the works of psychologist Stanley Milgram in the sixties [3]. Based on his experiment the author states that in average there is six degrees of separation that connect two randomly selected individuals in united stated of America. This surprising hypothesis, was studied in social networks. Particularly, Twitter was explored to investigate connectivity, the above hypothesis was confirmed. For instance, the data analytics company Sysomos [4] found out that on average 50% of Twitter members are four steps from each other, while everyone is five steps away (see figure 1.2).

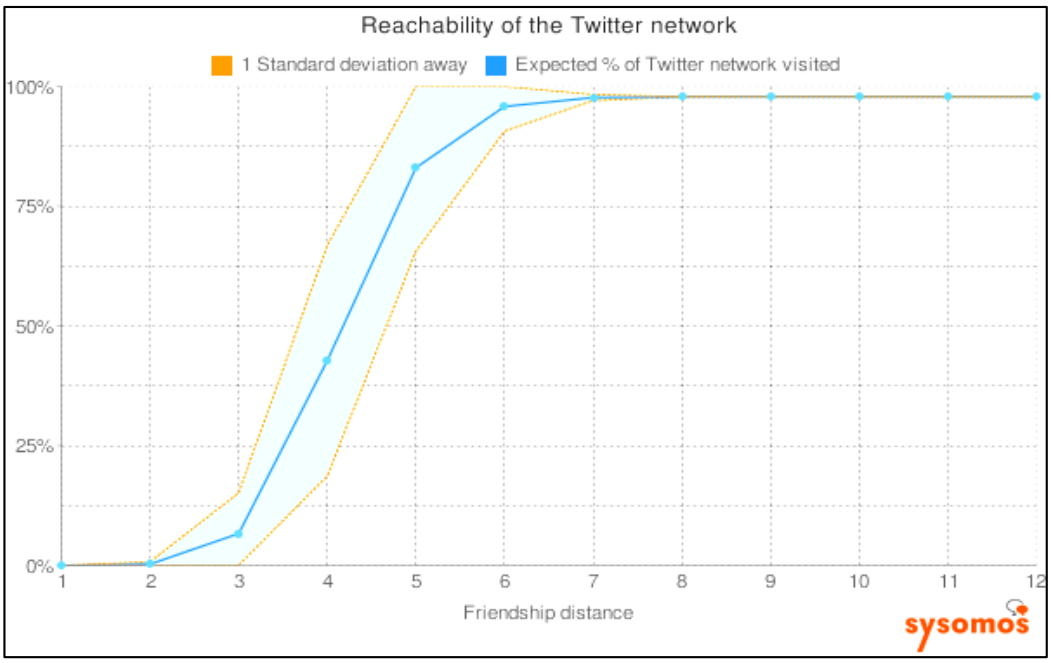


Figure 1.2 Illustration of friendship distance in Twitter (image from [4]).

I.3.2 Strength of weak ties

In his works Mark S. Granovetter [5] studied the strength of interpersonal ties or relation intensity in social networks to show the impact of this aspect relatively to information diffusion, social mobility and social cohesion in social networks. The author suggested that new information or innovation comes from weak ties. These findings could be explained by the fact that members with strong relations in a social network tend to form cohesive groups often called communities. Moreover, these social groups share basically common interests and same topics. Therefore, it is more likely that new information come from members outside the community or on its borderline.

I.3.3 Triadic closure

This principle of network transitivity is referred to as triadic closure, it expresses the fact that as the number of common neighbors between two nodes grows, they tend to be friends. As depicted in figure 1.3 below, if node A is a friend of node B and node B is friend of node C, it is likely that nodes A and C will become friends in near future. Newman [6] has computed this quantity in the context of collaboration networks, his findings show a positive correlation between the number of common neighbors of A and C at time t. Triadic closure is measured using a metric called the clustering coefficient of the network [7,8]. It is the average of the Clustering coefficient of all nodes that is defined as the ratio of the number of its connected neighbors to the number of all possible connection between pairs of neighbors of this node.

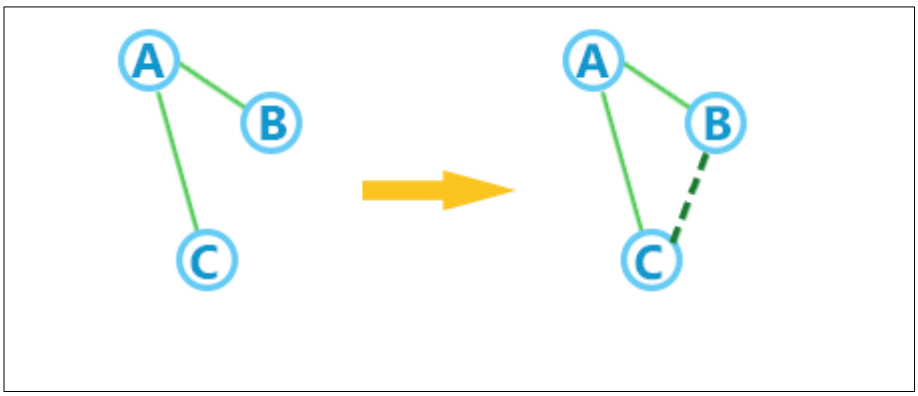


Figure 1.3 Illustration of triadic closure principle.

It expresses the probability that two incident edges close a triangle. The clustering coefficient measure ranges from 0 to 1. A value 1 means that all nodes are connected to each other to form a clique, (see figure 1.4) however a value of 0 indicates that there is no common neighbors connections. Formally, it is defined as follows:

$$Clustering\ coefficient(i) = \frac{2* n_i}{d_i\left(d_i - 1\right)} \quad (2)$$

n_i is the number of links between i' s neighbors and d_i is the degree of node i.

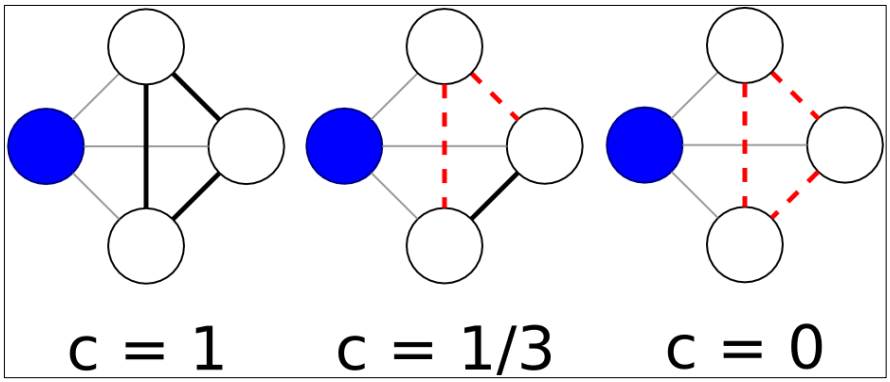


Figure 1.4 Illustration of clustering coefficient example**.**

I.3.4 Information diffusion

Information diffusion is the process that studies how information is spread in online social networks. It has gained an important impact on society; this could be seen through the role of Facebook during the Arab spring in 2010. The tremendous amount of content disseminated through these networks has attracted the attention of several actors ranging from researchers, companies to governmental agencies. Effectively, this content is being analyzed to the purpose of information extraction and prediction.

Moreover, several issues were investigated in this domain [9] such us, events detection, topics detection, influencers detection, etc. As illustrated in figure 1.4 below, Information diffusion is categorized in four types [10], according to that the network is implicit or explicit and considering if the information available to user is local or global.

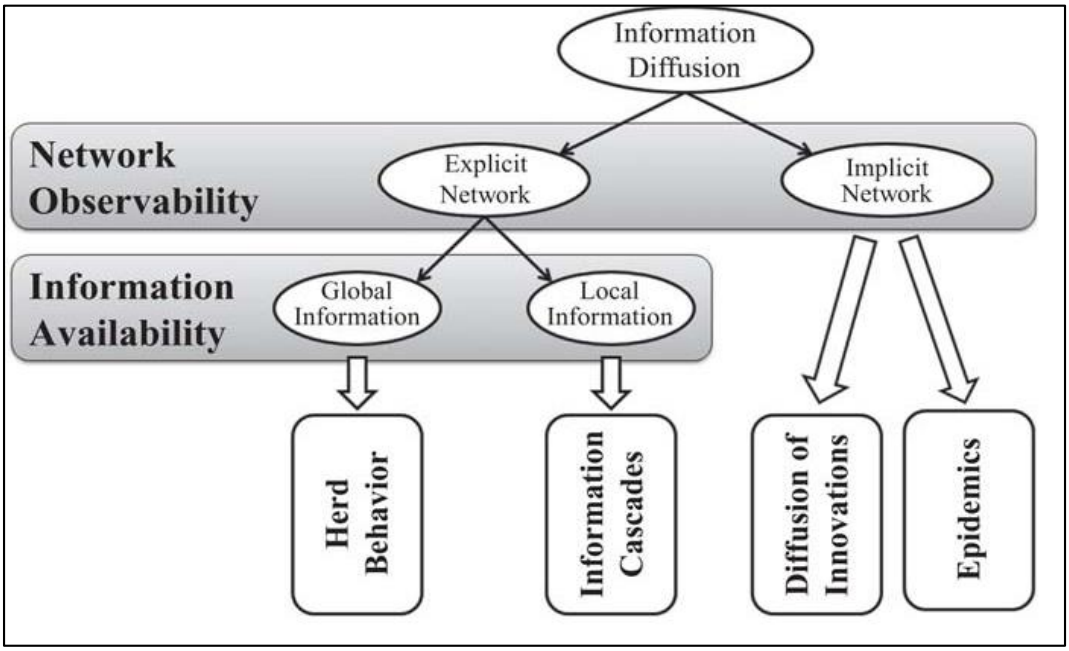


Figure 1.5. Information diffusion types in social networks (image from [10]).

In the cases of Herd behavior and information cascade the network is explicit which means for example that members participate in an online auction site, or they are linked and exhibit a relationship such as follower/followee in Twitter. In the other hand, diffusion of innovations and epidemics suppose that the interactions among people are not observable.

Herd behavior occurs when members observe all the actions of other and act as these later do (example online auctions), whereas in diffusion of innovation the only available information is the rate at which products are being adopted. Epidemics is similar to diffusion of innovation the only difference is that people do not decide to be infected.

I.3.4.1 Information cascade

Information cascade is observed if people ignore their own information signals and based on the actions of others, they adopt the same behavior [11]. To distinguish this process from the three other types of information diffusion, it should be noticed that in this case information available to users is local, i.e. they observe only the signals of their immediate neighbors, in addition to the fact that the network is explicit. As real-life examples in online social networks information cascade occurs when:

- Users of Facebook replicate a friends' post on their own wall.
- Followers in Twitter retweet a users' tweets.

- A user may not retweet the same message, but propagate piece of information related to the topic of its immediate neighbors.

Figure 1.6 illustrates an example of information cascade in social networks. It depicts pieces of information being cascaded on a path of friends in a network.

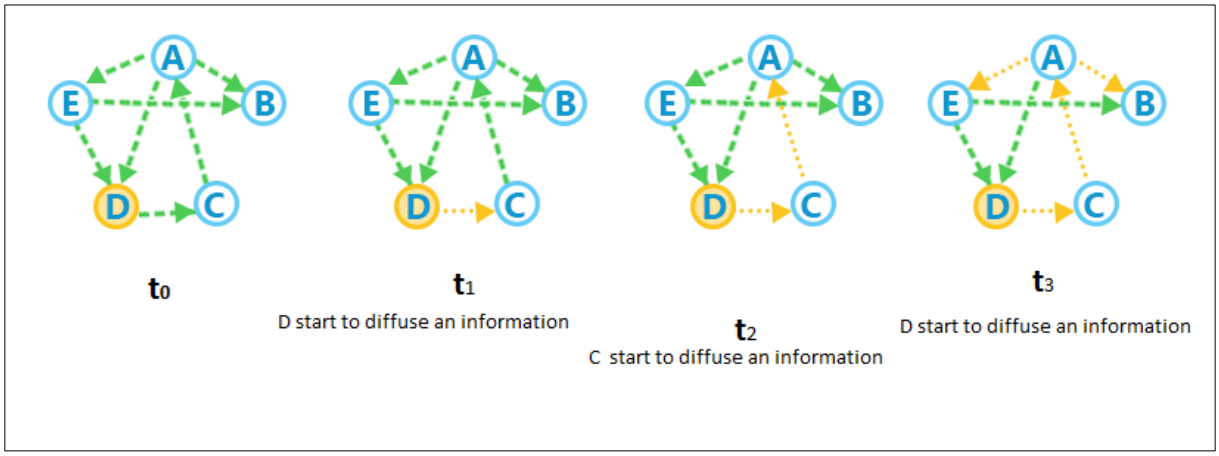


Figure 1.6 Information cascade illustration.

It should be pointed out that social influence is omnipresent in social networks. This notion was studied in the literature [9], its applications are various; for instance companies are interested in detecting influencers in online social networks to adopt a targeted marketing campaign, politicians show interest to this phenomenon for a purpose of propaganda, etc.

I.3.5 Communities as building blocks of social networks

In social networks members tend to form cohesive groups often called communities. Basically, these tightly knight members share same interests and topics. Formally, in a social graph $G = \{V, E\}$ a community C is a subset of nodes of V. The graph G could be partitioned in disjoint communities where each node should belong to only one community, whereas in the case of overlapping communities a node could be affected to multiple communities. This notion should be pointed out since it is an interesting level of granularity of social networks. Several researchers targeted communities detection in social networks which aim at portioning the graph in a set of communities, to the purpose of multiple applications such as: network evolution [12], recommendation system [13], and expert finding [14]. Studying communities leads to gain insights about members behavior as social groups. Moreover, information diffusion is another issue were it is important to consider the notion of community and its effect on the spread of information in social networks [15].

I.4 Conclusion

Online social networks play an important role in society, this chapter was dedicated to their introduction as well as some of their core concepts. Specifically, we have mentioned some of social networks interesting properties that have emerged from studies and experimental works in several disciplines. Moreover, we have highlighted the challenging issue of information diffusion in online social networks and its different types, given that it is tightly related to social influence. Finally, we have briefly discussed the importance of the notion of community in social networks and its interesting real-life applications.

II.1 Introduction

The fast evolution of the social networks such as Twitter, Facebook represent an important factor in social influence emergence. Influencers are a complex force in the internet and are considered as propaganda agents in the virtual world. Discovery of influential members could be the answer of who is the most important to follow?, or who is the most active member in the social network?, this is important in various domains such as viral marketing and political campaigns. Many recent research studies the influence detection in the last few years, much work in social network analysis has focused on the identification of influencers. In this chapter we will elucidate the detection of influence and its potential applications. Afterwards, some recent works in this domain will be presented.

II.2 Influencers detection and Applications

Social influence is the variation in behavior that one-person causes in another, in other words it is the change in a person's behavior, cognition, attitude, feeling, or sentiment, which has its origin in another person or group. This is the result of the way that changed person perceives themselves in relationship to the influencer that results from interaction between them**.** The detection of influencers is a recent line of research in social network analysis that has drawn the attention of many researchers given the exponential growing of available data on interpersonal relationship.

Identifying who is influential can help when planning political campaigns, advertising strategies, marketing and intelligent trade or even combating terrorism, influencer detection has allowed :

- Various companies to look beyond direct marketing to find potential customers to target, where these companies need to find the individuals who will spread more effectively information about the new products in which the influencer is the goal of these companies.

- Intelligent marketing decisions making which focus on the influencer marketing efforts to maximize viral effect of the group.

- Targeted political parties campaigns, for example for elections purpose.

- Successful advertisement strategies and recommendations.

- Detecting terrorist leaders, and promoting grassroots movements.

II.3 influencers detection state of art

Social influence has been a widely accepted phenomenon in social networks for many decades, in recent years researchers have developed a variety of techniques and models to capture information diffusion in online social networks, evaluate it, predict it, and extract knowledge data to discover spreaders in such domain. Two main methods for influencers detection were identified In the literature. The first type of methods employs the structure of the network for this purpose, whereas the second category is based on the content of interactions.

II.3.1 Structure based methods

These methods completely ignoring the content characteristics, use centrality measures, user's status and information diffusion approaches, to define and ranking influencers. In [16] authors have used a number of applications dependent assumptions about the weights of the different edges in certain social network. In some other works [17, 18] authors have identified influencers based on the notion of centrality. In [19] Kiss et al. have detected influential nodes that are more centrally connected in the social network. In Twitter many influence measures include number of followers, mentions, and retweets. For instance Authors in [20,21] have used Pagerank of follower network.

Moreover, Wei Chen et al. [22] have discussed the problem of information diffusion in networks and have used the diffusion model to find the top influencers. Their method is based on a fixed and static view of the network in which the edge diffusion probabilities are assumed a-priori and do not focus on characterizing the cascading behavior from the flow of information. Other work that treats the influencers mining problem in a supervised classification setting, in [23,24] authors have addressed the problematic of detecting influence by posing it as a predictive task.

Many different models proposed to capture and/or predicting the diffusion process in the network, which is characterized by its structure and its temporal dynamics. The process of modeling information diffusion developed in two categories of models, on the one hand there are explanatory models, objective is to infer the underlying spreading cascade, which are very useful to understand how information propagated, for example it is possible to take the path taken by a piece of information in such static and dynamic network. Authors in [25] have suggested to discover the correlations in nodes infections times to infer the structure of the

spreading cascade and assume that activated nodes influence each of their neighbors independently with some probability that one node had transmitted information to another.

On the other hand predictive models, which aim at predicting how a specific diffusion process describe in a particular network as of temporal and/or spatial points of view by learning from historical information diffusion traces. In [26] the main idea is the use of static edge broadcast probabilities, the problem seeks for finding a small set of seed nodes in a social network that exhibit most influence spread in certain influence cascade models, such as linear threshold cascade or independent cascade models. In [27] authors have proposed a partial differential equation based model to predict the diffusion of an information introduced in the network by a given node.

II.3.2 Content based methods

These methods capture the dynamic nature of the content flows. works in [28] consider the potential importance of topics and constructs a weighted graph indicating similarity, in terms of topics, between two users. The traditional detecting interesting topic techniques developed to focus on burst, Leskovec et al. [29] have developed scalable algorithms for clustering textual variants of such words, their work offers some of the first qualitative analyses of the dynamics of information propagation. Shamma et al. [30] have developed a method that counts the number of tweets that a term appears in, rather than raw total frequency of the topic, they developed two methods of finding temporal topics with stream of Twitter using a normalized topic frequency.

To identify the influential spreaders in such network some methods illustrate the several possible techniques to measure the influence of each node in an online social network, for example k-shell decomposition [31] or HITS [32] to textual clustering based methods including hybrid methods.

Authors in [28] have proposed a similar measure to Page Rank to quantify influence on Twitter, and have developed a topic-centric version of the Twitter Rank algorithm designed especially for Twitter. In [33] authors have developed a non-graph based topic-centric method, they have defined a set of modal and topical features for characterizing the network members, to the purpose to identify the most influential for a given topic, they rank nodes according to a within-cluster ranking method.

In [34] the proposed approach consists of two main phases; first, they represent user-generated content in a high-dimensional space so that an order of user-generated posts is mapped to a time-series in this space. Secondly, they apply information-theoretic measures to those time series to determine and quantify directed influence among the users. Because their method is based on information theoretic principles, it is easy to interpret, and flexible with respect to the representation of content.

In 2013, A. Guille et al. [9] have proposed a taxonomy illustrated in figure 2.1 to define the three main research challenges arising from information diffusion in online social networks.

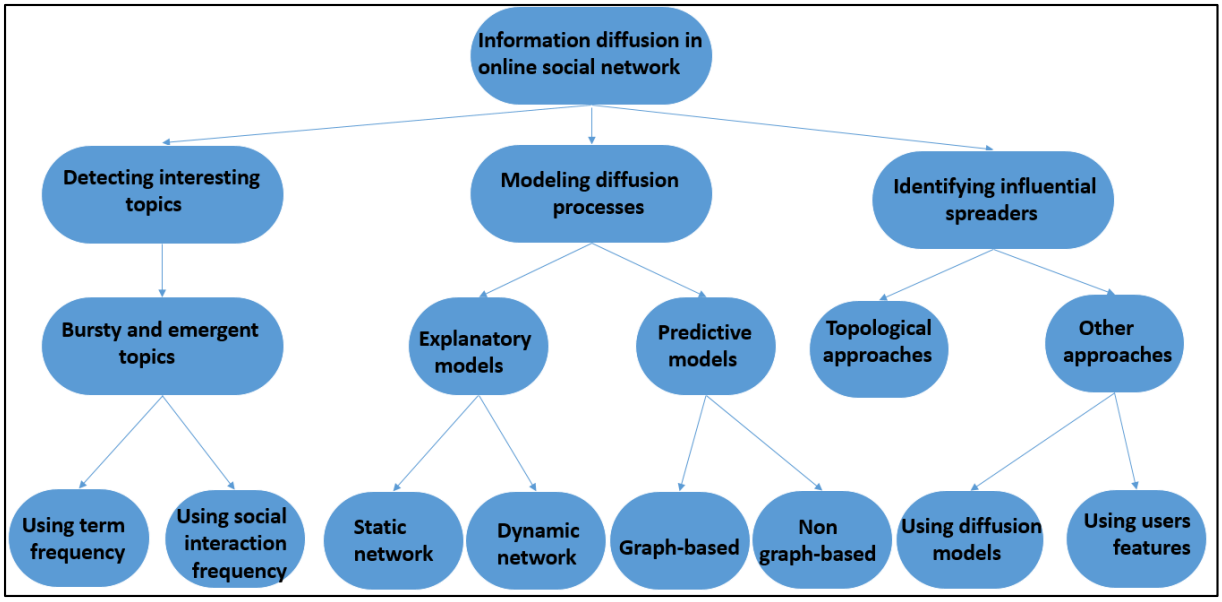


Figure 2.1 Main research challenges arising from information diffusion in online social [9].

It should be pointed out that since the works in [9] new approaches were proposed to contribute in the identification of influential spreaders. Some of these methods are based on the content of interactions as well as the underlying network.

To update the taxonomy mentioned above, One should consider the methods using the propagation of content through the social graph. For instance, we could cite the works of S. Karthik [35] that should be considered in this taxonomy. Figure 2.2 below illustrate the proposed update.

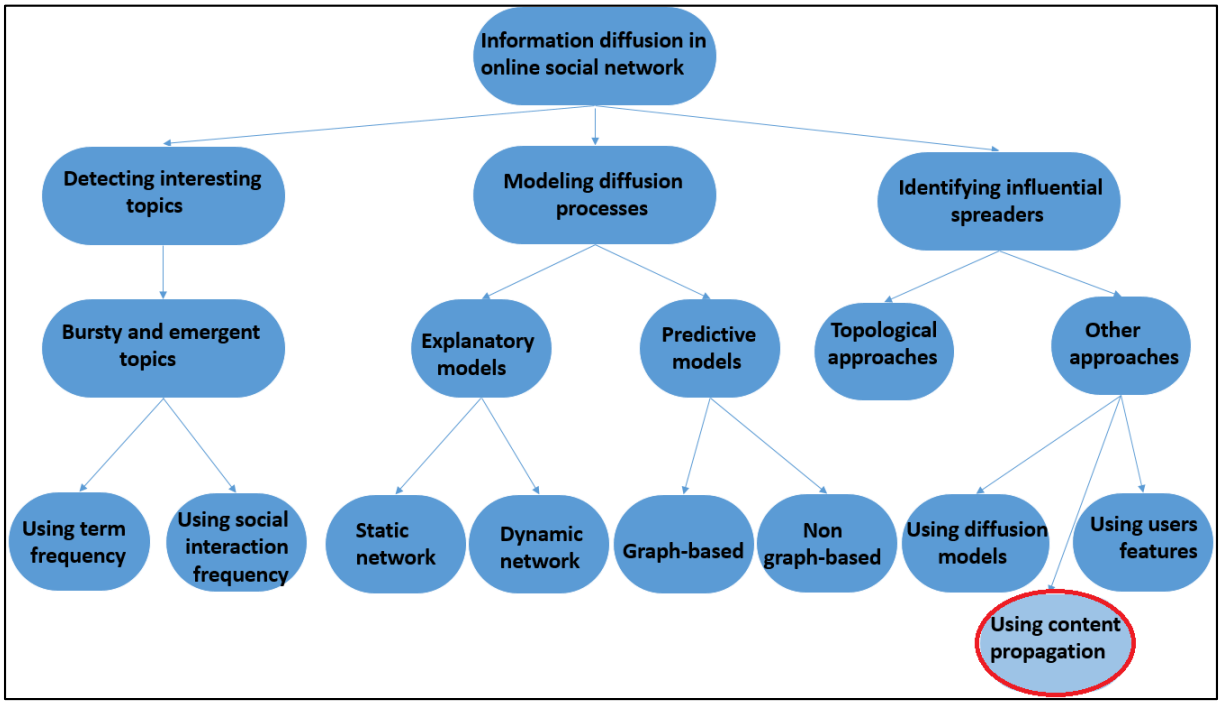


Figure 2.2 Updated taxonomy of information diffusion.

II.3.2.1 Information flow mining algorithm

In [35] and later in [36] authors have studied the problem of influence detection in the context of information flow in networks. They have designed an algorithm to detect information flow patterns in social networks. Specifically, they first targeted the detection of paths (set of linked nodes) through which exact information (text) content is propagated more frequently, they refer to these paths as frequent paths, considering in that the chronological order of users information posting. Afterwards, they computed a score using the mentioned frequent paths to discover influencers in social and collaborative networks.

The content generated by the different users of the network was considered in this work as a social stream of text content, each element of this social stream represents a unit of information called content-token, and its originating actor. Formally, a social stream is defined as all couples (U_j, a_i) where U_j is the posted token and a_i is the originating user.

Definition 1 (Valid flow path)[35]. Let $G(N, E)$ be a network, a valid flow path is an ordered sequence of distinct nodes $n_1 n_2 \dots n_k, n_i \in N$, such that for each t ranging from 1 to $k-1$ an edge exists between nodes n_t and n_{t+1}.

Definition 2 (Information flow frequency)[35]. The information flow frequency of actors $n_1 n_2 \dots n_k$ is the number f of units of information content (tokens) $U_1 \dots U_f$ given that for each U_i the following conditions hold:

- Each actor from the sequence $n_1 n_2 \dots n_k$ has posted U_i
- Each U_i was posted by the actors in the order of the sequence $n_1 n_2 \dots n_k$.

Definition 3 (Frequent path) [35]. A sequence of actors $n_1 n_2 \dots n_k$ is defined as a frequent path with frequency f, if the following two conditions hold:

- The sequence of actors $n_1 n_2 \dots n_k$is a valid flow path.
- The information flow frequency of the actor sequence $n_1 n_2 \dots n_k$ is at least f.

Figure 2.3 below illustrates the notions introduced in the above definitions. The problem of information flow patterns mining was defined in [35] as the discovery of all frequent paths of a given frequency f, given a set of tokens $\{U_1, \dots, U_n\}$ of a social stream and a social graph G.

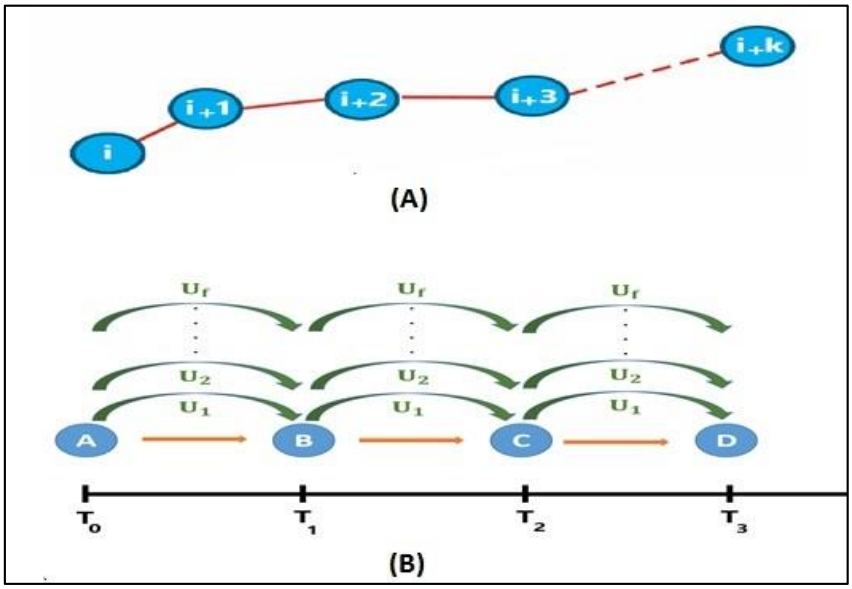


Figure 2.3 Illustration of the above definition **(**A) example of valid flow path (B) example of frequent flow path with frequency=f**.**

The algorithm *InMineFlow* [35] described below aims at detecting all frequent paths, i.e. valid paths with a frequency of at least f. In what follows $\mathcal{F}_{a_i}^k$ stands for actor a_i (a_i being the first actor of the sequence) frequent paths of length k. The algorithm starts generating all singleton frequent paths, they represent all paths composed each of one actor who have at least f posts in the social stream. The next step of the algorithm seeks for candidate frequent paths of length

2 noted $\mathcal{P}_{a_i}^2$ of each actor a_i . This set of candidates of length 2 is obtained using $\mathcal{F}_{a_i}^1$ such that all frequent neighbors of actor a_i are added to paths of $\mathcal{F}_{a_i}^1$. Afterwards, a pruning step is performed based on the fact that all k-length paths of $k+1$-length frequent paths are frequent paths. Hence, the candidate frequent paths for which at least one k-length path is not frequent are not considered. Then the algorithm tests if the remaining candidates have a frequency f and finally generates $\mathcal{F}_{a_i}^{k+1}$. The algorithm iterates until there are no frequent paths to generate.

Algorithm *InMineFlow* (Social Stream: *T*, frequency: f);

Begin

for each node i

$\mathcal{F}_{a_i}^1 = \{\}$;

if node i has at least f posts in T then

$\mathcal{F}_{a_i}^1 = \{a_i\}$;

end

$k = 1$;

while $\cup_i \mathcal{F}_{a_i}^k$ is not empty do

for each i generate $\mathcal{P}_{a_i}^{k+1}$ from $\mathcal{F}_{a_i}^k$ by appending each frequent neighbor node of i to the paths in $\mathcal{F}_{a_i}^k$, and keeping only potential paths with all distinct nodes;

for each i prune any path from $\mathcal{F}_{a_i}^{k+1}$, for which at least one of its path of length k is not in $\mathcal{F}_{a_r}^k$ (for some r);

for each i generate the path in $\mathcal{F}_{a_i}^{k+1}$ which are valid frequent flow paths;

Reorganize the frequent potential paths $\cup_i \mathcal{P}_{a_i}^{k+1}$ based on the last nodes of these paths, and generate the corresponding path lists $\mathcal{F}_{a_i}^{k+1}$ for each node i ;

$k = k + 1$;

end

return $(S(f) = \cup_k \cup_i \mathcal{F}_{a_i}^k)$

end

To illustrate this algorithm, if we consider the social stream T and the frequency $f = 3$ where

T= {(U_1, A) (U_1, B) (U_1, G), (U_2, A) (U_2, B) (U_2, G),(U_3, A) (U_3, E) (U_3, D) (U_3, X), (U_4, A) (U_4, E) (U_4, D) (U_4, X) , (U_5, A) (U_5, E) (U_5, D) (U_5, X) (U_5, G) , (U_6, D) (U_6, B) }

And the social graph shown in figure 2.4.

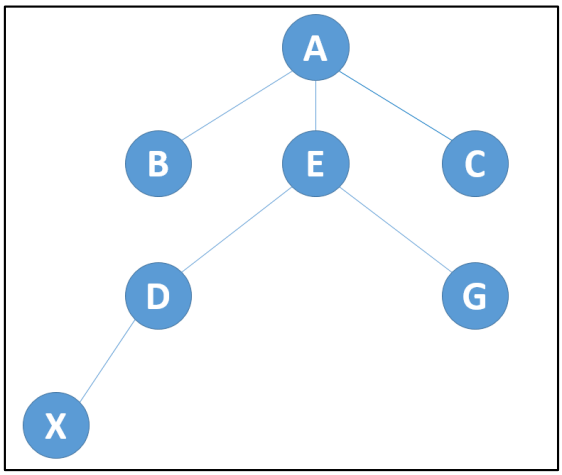


Figure 2.4 Illustrate social graph example**.**

The algorithm proceeds as follows. It starts generating all singleton frequent paths,

$$\mathcal{F}_A^1 = \{A\}, \mathcal{F}_B^1 = \{B\}, \mathcal{F}_C^1 = \{C\}, \mathcal{F}_D^1 = \{D\}, \mathcal{F}_E^1 = \{E\}, \mathcal{F}_X^1 = \{X\}, \mathcal{F}_G^1 = \{G\}.$$

Next candidate sets of size 2 will be generated by appending different neighbors from the social network. The resulting frequent paths candidate set $\mathcal{P}_i^2$ are:

$$\mathcal{P}_A^2 = \{(\mathrm{A}, \mathrm{B}), (\mathrm{A}, \mathrm{C}), (\mathrm{A}, \mathrm{E})\}$$

$$\mathcal{P}_B^2 = \{(B, A)\}$$

$$\mathcal{P}_C^2 = \{(C, A)\}$$

$$\mathcal{P}_D^2 = \{(D, X)\}$$

$$\mathcal{P}_E^2 = \{(E, G), (E, D)\}$$

$$\mathcal{P}_X^2 = \{(X, D)\}$$

$$\mathcal{P}_G^2 = \{(G, E)\}$$

Prune candidates:

$$\mathcal{F}_A^2 = \{(\mathrm{A}, \mathrm{E})\}$$

$$\mathcal{F}_B^2 = \{\emptyset\}$$

$$\mathcal{F}_C^2 = \{\emptyset\}$$

$$\mathcal{F}_D^2 = \{(D, X)\}$$

$$\mathcal{F}_E^2 = \{(E, D)\}$$

$$\mathcal{F}_X^2 = \{\emptyset\}$$

$$\mathcal{F}_G^2 = \{\emptyset\}$$

This step can be reiterated in general from cascades of k-length to (k + 1)-length ones, and then further pruning and support counting. Now, we will find candidate sets of size 3 using $\mathcal{F}_i^2$ by appending frequent neighbors, after the second iteration of *InMineFlow* as described below.

$$\mathcal{P}_A^3 = \{(\mathrm{A}, \mathrm{E}, \mathrm{G}), (\mathrm{A}, \mathrm{E}, \mathrm{D})\}$$

$$\mathcal{P}_E^3 = \{(E, D, X)\}$$

Prune candidates :

$$\mathcal{F}_A^3 = \{(\mathrm{A}, \mathrm{E}, \mathrm{D})\}$$

$$\mathcal{F}_E^3 = \{(E, D, X)\}$$

When $k = 3$ append the candidate set:

$$\mathcal{P}_A^4 = \{(\mathrm{A},\mathrm{E},\mathrm{D},\mathrm{X})\}$$

$$\mathcal{P}_E^4 = \{\emptyset\}$$

Prune candidates

$$\mathcal{F}_A^4 = \{(\mathrm{A}, \mathrm{E}, \mathrm{D}, \mathrm{X})\}$$

Next iteration

$$\mathcal{P}_A^5 = \{\emptyset\}$$

Now the algorithm stops as there is no further candidates to process. The final result is the set $S = \{(A), (B), (C), (D), (E), (G), (A, E), (E, D), (D, X), (A, E, D), (E, D, X), (A, E, D, X)\}$.

II.3.2.2 influence of an actor

To quantify the influence of an actor the authors in [35] have considered two main factors namely the position of the actor in the frequent flow path, and the number of frequent paths that involve this actor. Given a social graph $G(N, E)$, a set of frequent paths $\{p_i\}_{i=1}^n$ in which an actor a_k occurs, let $Pos(a_k, p_i)$ be the position of actor a_k in the frequent path p_i. the influence of a_k is defined as follows:

$$Influence(a_k) = \sum_{i=1}^{n} |\{a_r \in p_i, such\ that\ Pos(a_r, p_i) > Pos(a_i, p_i)\}| \qquad (3)$$

Intuitively, the influence of an actor a_k is the number of actors that appear after a_k in all the frequent paths where a_k occurs.

II.4 Conclusion

In this chapter, a state of art of influencers detection in networks was presented. Particularly, we have focused on the *InMineFlow* algorithm, which represents a relevant solution for information flow patterns discovery. This later targeted the diffusion of the exact content of information (content-tokens) to detect frequent flow paths, which were used to quantify the influence of actors using both of the social stream and the social graph. The next chapter is dedicated to our proposed approach, which is mainly based on this algorithm

III.1 Introduction

In the state of art we have presented methods for influencers detection based on information cascade such as *InMineFlow* algorithm. This algorithm as discussed previously targeted the propagation of exact content to discover influencers. The action of retweeting a post in Twitter is an example of the propagation of the exact content, this could be seen as strong social influence. So we asked if a non-direct use of the exact unique content in the mining process would affect the result or not?

In this chapter we will further describe this proposed idea that could lead to an improvement of the previous methods since it seeks for influencers detection based on both exact content-tokens propagation and their related keywords cascade.

III.2 Problem formulation

Suppose a user A on a social network starts the diffusion of information containing a set of tokens T: T_1, T_2, ... T_n . When another user X alternates those pieces of information by changing the syntax, or personalizes the content using other tokens in a way that conserves the same meaning of T. The detection of frequent paths ends without containing X and the users that follow him. Clearly, if the first diffusion of a specific token, let us say "USA", after that someone somewhere personalizes it to "America" or "United states of America", *InMineFlow* will not detect this last one and the reason is that it only focuses on the token just as it is.

Moreover, if our supposedly first actor A had already personalized this set of information after being influenced by another actor $alpha$, then the actor that originated that information was not A but $alpha$.

III.3 Proposed approach

Before we explain our approach, let's first put this problematic in generalized words, If there is a token T and a set of keywords K where $K = \{ M'\ M''\ M'''\ M''''\}$.$\forall\ \mathrm{M}^i \subset K$, M^i is a keywords related to T.

The reasonable solution is to account for all these "possibilities", that would only be feasible if we upgrade the tracking from the syntax based way to the semantic one. Therefore, we propose to track not only the same tokens or the words, but also consider their related keywords to gain some intuition about the detected cascades. In simpler words, we track T as

it is and any of others keywords related to it in a fashion that guarantees that we conserve the cascade, no matter if one or many other actors personalized the information.

The first step of our approach is to detect frequent paths and collect the involved content-tokens from the social network using the content and the social graph. This step is performed by mean of our algorithm *PaToDet*. Afterwards, we leverage Web intelligence to extract related keywords to the content-tokens. The work done by A. Singhal in his PhD thesis [37] "Leveraging open source web resources to improve retrieval of low text content items" was our source of inspiration. The author used Google search API and Alchemy API to the purpose of low content text tagging.

This inspired idea must be shaped into a simple clear method, a method that enables us to generate a list of keywords by only inputting one (or composed) keyword (keywords). For this purpose we have used GOOGLE search API to extend the collected content-tokens context, and ALCHEMY API to extract keywords from this extended context.

The main steps of our approach are:

- Apply *PatoDet* algorithm to get the frequent paths and collect the tokens involved in these frequent paths.
- Leverage Web intelligence to extend the context of each involved token to extract relevant keywords related to it using the following tools:
  - Google APIs [38].
  - AlchemyAPIs [39].
- Apply *PatoDet* again on the tokens and their new associated keywords to mine the new frequent paths.
- Compute the scores of each node of the social graph.

III.3.1 Major Steps

Let us first define, before listing the major steps, those techniques used for leveraging web intelligence to extract keywords.

III.3.1.1 Google APIs

Google APIs represent a collection of applications programming interfaces, developed by Google, in Order to permit a communication with Google services and also integrate them in

other services. For example Research, Gmail, Google Map and others. The APIs deliver as well a set of functionalities such as analytics, the PredictionAPIs (from deep learning ,machine learning) and if the permission to read and write the data is given another service can be allowed as access to user data [38].

A simple protocol is used in some of APIs witch requiring an authentication and authorization, one need first to get identification information's from the developers Console, after that the client application required an access key delivered by Authorization Server of Google, finally this key is used when accessing to a Google APIs service. Client libraries are unveiled in multiple programming languages as JavaScript, Objective-C, .NET, PHP and Python to help developers to use Google APIs from their code.

III.3.1.2 ALCHEMY API

It is an IBM company that aims making easier the natural language processing, such as (semantic text analysis and the sentiment analysis) and computer vision (face detection and recognition) for customers in the cloud and on premises deep learning and machine learning.

AlchemyApi provides a suite of text analysis functions and picture. These functions are available through a REST AP, and the data is sent a simple structure to analyze what can be integrated into your application. AlchemyAPI's text analysis functions include:

- Extraction of the entities
- Keywords Extraction: extraction of the important terms, figure 3.1 gives an illustration of this function. In our approach we have used this function to get some relevant keywords, we give more details below.
- Extract the concept - identify the broad concepts of the text.
- Extract the Relation (subject-action-object relations).
- Classification: categorize automatically your text.
- Detect Language.



Figure 3.1 Alchemy Api keywords extraction illustration.

The major steps of the proposed approachare:

- **Step 1:** Apply *PaToDet* using the social graph and the social stream. This phase will generate all the frequent paths, and tokens that have walked through these paths, which will be considered as keywords that will be used in the following step2).
- **Step 2:** Extract relevant keywords using the web intelligence**.** To apply this step we have used two functions first, to query GoogleApi search engine and the second to query AlchemyApi . This steps will be divided into two sub-steps:
    - Extracting K-top titles and snippets (*googleAPI*)

      After gathering the keywords from the previous step 1, we input each content-token individually in GOOGLE-API (the content-tokens might be composed). The output results will not yet be final, but will be more refined by limiting K to a certain value, then aligning those titles and snippets in a coherent paragraph-like fashion.
    - K most relevant keywords (*AlchemyApi*)

      Results gathered from the previous step, will be transferred to ALCHEMY-API, which via the "***Keyword extraction***" and "**Concepts Extraction**" functions will treat the

aligned tokens and extract a set of relevant keywords sorted by a relevance score (then eliminate keywords with scores under ~70%).

We remind that the efficiency of *PaToDet* algorithm (see algorithms 1 and 2 below) is the way that it simplifies the string matching and frequency tracking process, it uses a hash table as in [36] to track the sequence of actors for each distinct content-token in the stream. In this step, we re-organize the social stream content by adding the relevant keywords from the step two, while using the same hash function. We add the relevant keywords always using the same hash function, to conserve the time order, the list of actors of each relevant keywords will be related to the original keyword in the hash table.

Step 3: Apply *PaToDet* again on the hash table described in section 3.4 which contains the list of the frequent tokens and their relevant keywords obtained from step 2.

Step 4: Compute the scores of each node as explained in section 3.2.2 of chapter 2 to detect the influencers (i.e. nodes with the highest score).

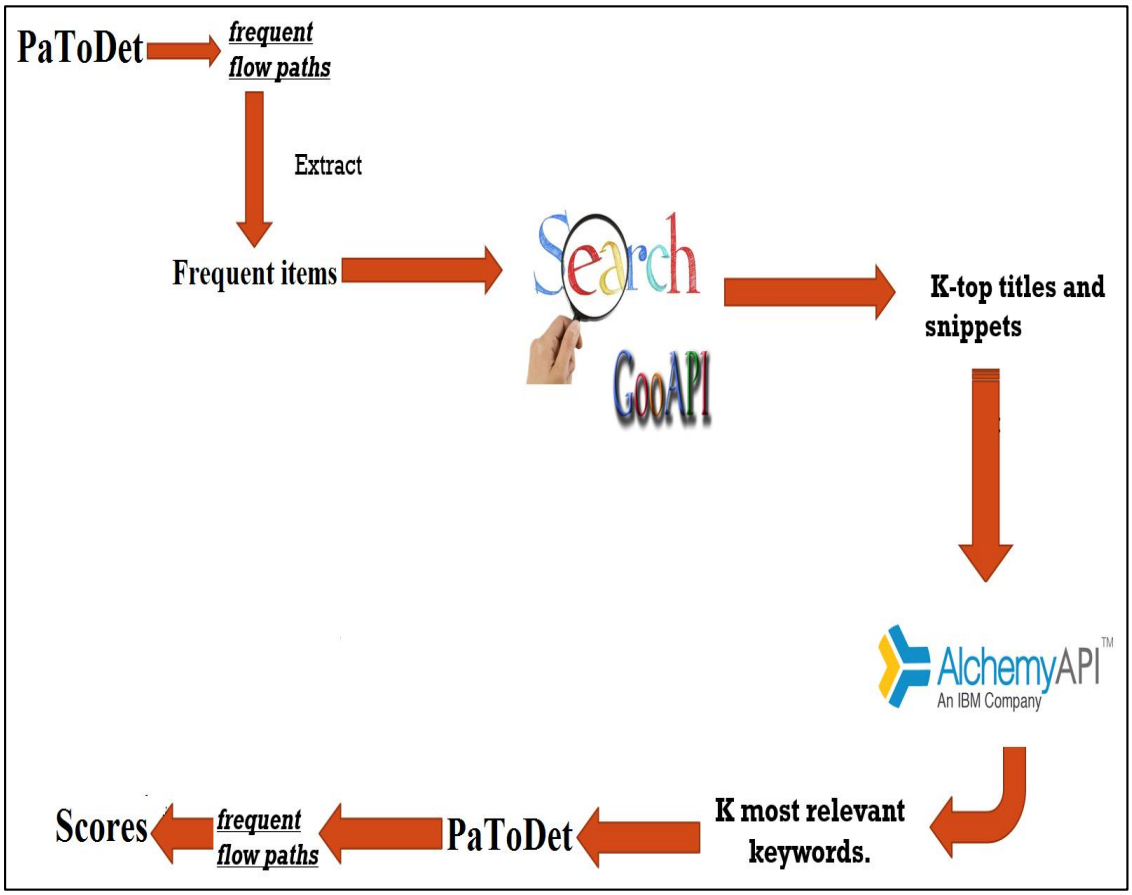


Figure 3.2 Approach architecture.

Algorithm 1. *PaToDet ((*Social Stream: T, frequency: f , frequent item: ft);

Begin

Generate all f token from ft list

For each T_i in tokens

Find all frequent actors who posted T_i ;

added frequent actors ;

end

$list = \cap_f$(actor list);

While $list$ is not empty

For each a_i in $list$

If (frequent neighbors of $a_i \in list$)

Save tokens

$a_i.score = n$; //n is number of a_i frequent neighbors in $list$.

Add node score (a_i, a_i.score);

Add edge list (a_i,a_j) ; //a_j is the frequent neighbors in $list$;

end

end

end

end

Algorithm 2. *CalculateScore*(Edgelist :El ; nodescore: Ns);

Begin

For each a_i in Ns

$S = 0$;

Return Score (a_i, S, El);

end

end

III.4 Experiments

To the purpose of validation of our approach, and given that dataset of Twitter are no longer sharable, we assimilate DBLP to a social Network as in [36]; the titles of co-authored papers were considered as posts and the graph of co-authorship as Follower/Followee network.

Before we elucidate this method's steps, we should shed the light on the DBLP dataset that we used as a trial data (which was also used in [36]) and the principal preprocessing phase to prepare our data file in order to test the proposed approach.

DBLP dataset is available as an XML file [39], it had 1707768 authors, 3308458 publications 4685 conferences and 1469 journals. In DBLP the relationships of co-authors make the underlying network available through their publications. The information shared through the co-authorship network could be, key research ideas, terms and technologies.

One author may initiate the research idea and may spread through their collaborators to different regions of the network. As in [36] we can create a graph from DBLP dataset where the authors are the nodes and the relation between them is the edge, we define the relation between authors by the act of writing a paper together: an edge exists between A and B if author Aco-writes with author B.

III.4.1 Dataset preprocessing

Once the dataset downloaded After we download as an XML file (see figure 3.3), we have performed the following actions:

- Extract the title, date and authors.
- Eliminate all information with messing data.
- Pre-process the title by treating every existing word as a unique content, applying stop words filter, non-alphabetic character removal, length 2 token removal and transforming characters from big to small letter.
- Codify the author's name, this procedure will reduce the complexity by minimizing the rate of comparison.

  We have used a function to codify authors as shown in this example, for a given actor 'Hans Ulrich Simon' its corresponding code is 'ha2267on'.

  This function is beneficial in terms of time costs, because it spares us of the need to verify whether the new author's code already exists or not.

- Create an adjacency list of the relationship between authors as follows: if A co-writes a paper with B, A also co-writes with C and C co-writes with B. with this set of information we create an adjacency list as illustrated below:

  A , B C

  B , A C

  C , A B

Given this adjacency list, we have created an undirected graph.

- The final step is sorting the data by date to conserve the time order (from oldest to the newest). Figure 3.3 is a snapshot of the social stream document, it illustrates the way data looks like.

Figure3.3 Snapshot of the original DBLP XML file.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE dblp SYSTEM "dblp.dtd">
<dblp>
<article mdate="2011-01-11" key="journals/acta/Saxena96">
<author>Sanjeev Saxena</author>
<title>Parallel Integer Sorting and Simulation Amongst CRCW Models.</title>
<pages>607-619</pages>
<year>1996</year>
<volume>33</volume>
<journal>Acta Inf.</journal>
<number>7</number>
<url>db/journals/acta/acta33.html#Saxena96</url>
<ee>http://dx.doi.org/10.1007/BF03036466</ee>
</article>
<article mdate="2011-01-11" key="journals/acta/Simon83">
<author>Hans-Ulrich Simon</author>
<title>Pattern Matching in Trees and Nets.</title>
<pages>227-248</pages>
<year>1983</year>
<volume>20</volume>
<journal>Acta Inf.</journal>
<url>db/journals/acta/acta20.html#Simon83</url>
<ee>http://dx.doi.org/10.1007/BF01257084</ee>
</article>
```

Figure 3.4 shows how the data was structured in the social stream data file, it is split into two ranges, a set of actors, where the first one sends to the second those sets of tokens in range two (the ranges are separated by a coma) , then sends to the third and so on …

va2735ko va1242ko , existence complete sets bounded reducibilities.
5939is , adaptive demodulation rateless erasure codes.
su955im ta931ng eu881uh hy1655ng , customer segmentation strategy development based customer lifetime case study.
we341en qi561ao an1630mi , power control cognitive radio networks cross multi-lane highway.
ho722im se1011oh ji1193im , smooth surface extension curvature bound.
ro1705ka , constructing fair curves surfaces sobolev gradient method.
jo1566rs xi865wu , sleves planar spline curves.
gu1850nn , degree elevation developable be zier surfaces.,
ar1053as va1195ro ig2259ov , heterogeneous material modeling distance fields.
qu1284an , efficient groebner walk conversion implicitization geometric objects.
jo1694ry , shape-preserving approximation multiscale univariate data cubic l1 spline fits.
se755uc bi53603 je2463en qu299mo , dyadic hermite interpolation square mesh.
ke2769as jo1566rs ul728if , shape characterization subdivision surfaces--case studies.

Figure 3.4 Snapshot of the social stream data file.

The Hash function: $h\ (U)$ is a hash function where U is a given token:

- We put the alphabets in order, then accord to each one a value ranging from 1 to 26 (in the same ascending order) to create an index table:

a➔ 1

b➔ 2

c➔ 3

:

y ➔ 25

z ➔ 26

For a given token U we multiply the position of each character in the token by its index value from the index table, afterwards we sum up all multiplication's results to get the "hash value" of U. for more clarification, let's consider the following example; let the token be the word "data", according to the index table, the letter "d" is indexed 4 the letter "a" is indexed 1 and so on …

$$h(data) = 1 * 4 + 2 * 1 + 3 * 20 + 4 * 1 = 70$$

It should be pointed that DBLP uses only scientific terms, so the use of the character's position and the index table will avoid any kind of collision.

The purpose behind re-applying *PaToDe*t is to enlarge the size and number of the involved actors resulting in the augmentation of detection's precision.

The closure step is to calculate the newest score of each node and compare it to the first result obtained in the step 1.

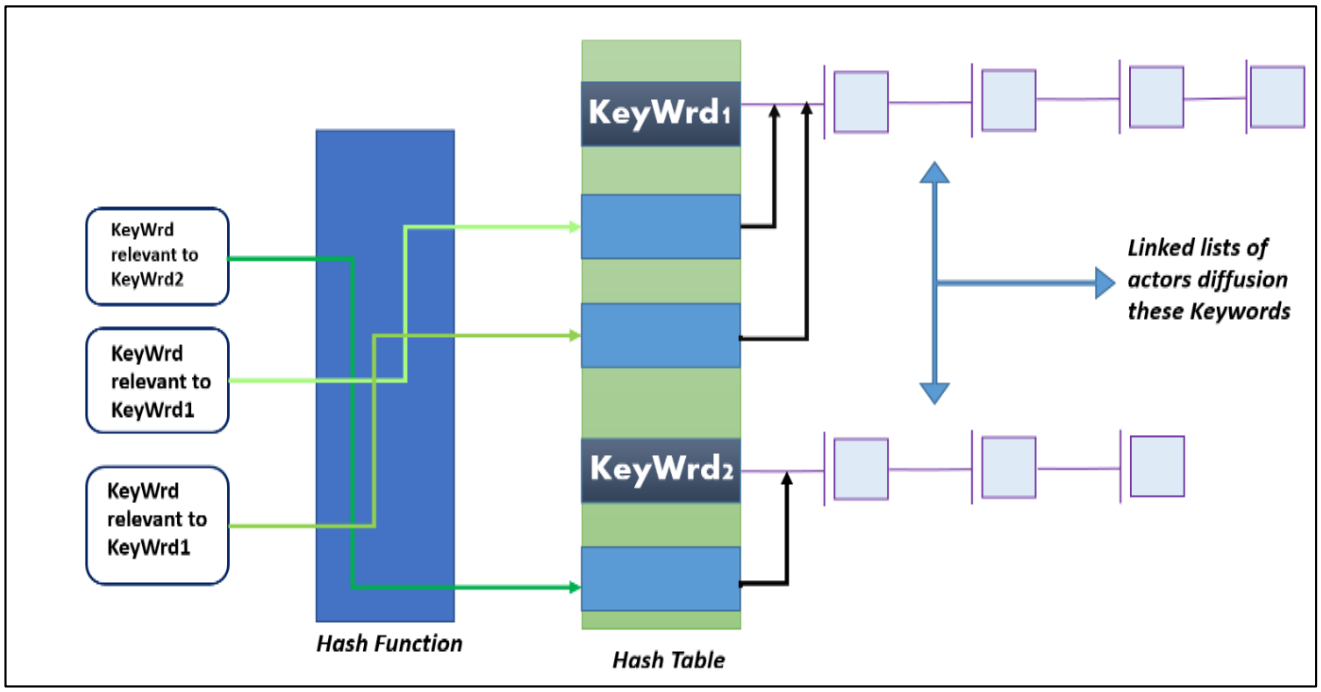


Figure 3.5 Illustration of the social stream reorganization for the third step in *Detectpaths2*.

III.5 Results

In this section, we will evaluate our proposed approach on a windows 64-bits machine, manufactured by MSI model (GE40 20C), with Intel i7-4702MQ CPU 2.20GHz processor, 8 GB RAM and hard disk of 320 GB.

Our dataset is composed of 2010 records of DBLP dataset, we have created the social stream as we have mentioned before, we have kept all the lines sorted to conserve the time order, each line contains two columns separated by a comma :

1st column: set of author's who have written same title.

2nd column: the same title after been cleaned.

After the first execution, we calculate the score of nodes and we pick the top 11 influencers as shown in figure 3.6.

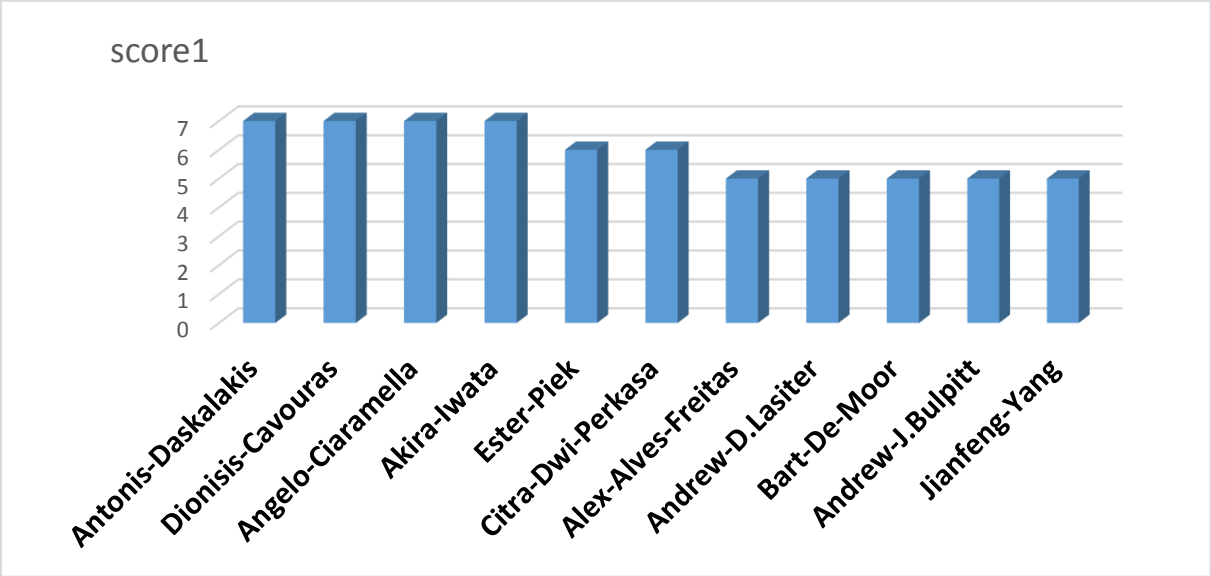


Figure 3.6 Top 11 influencers obtained from the first execution.

After the second execution of *PaToDet* with the relevant keywords obtained from AlchemyApi and GoogleApi, many changes were observed in figure 3.7 and figure 3.8. We have noticed that new actors show up as new top influencers which is due to the effects of adding the relevant keywords. Figure 3.7 illustrates the results of the second execution of *PaToDet* on the extended Hash table.

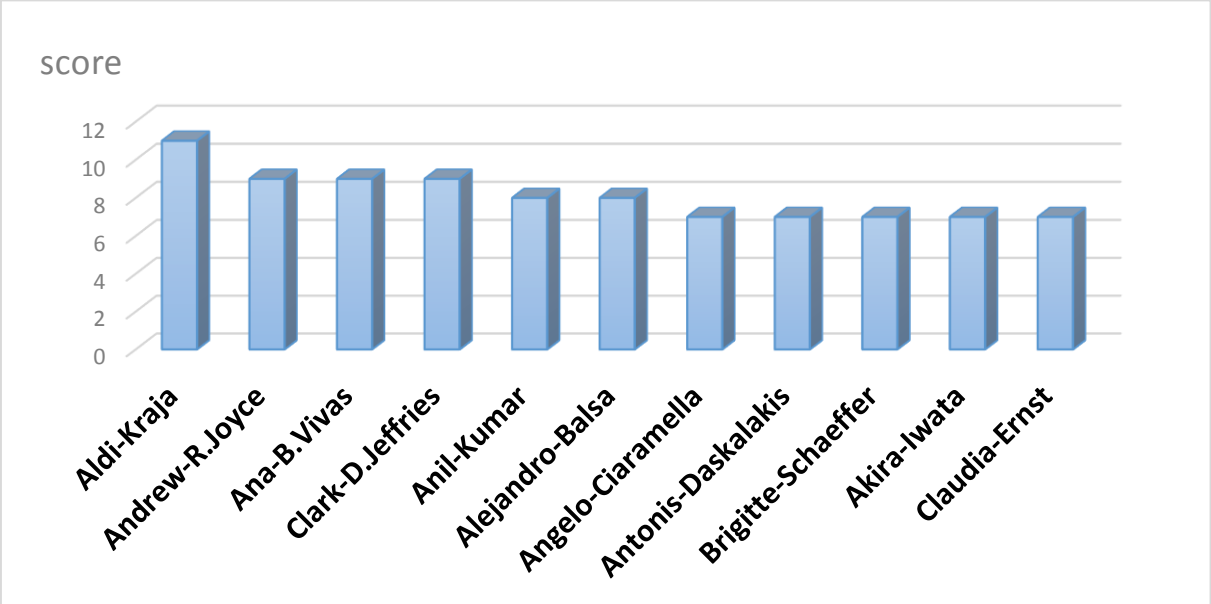


Figure 3.7 Top 11 influencers obtained from the second execution.

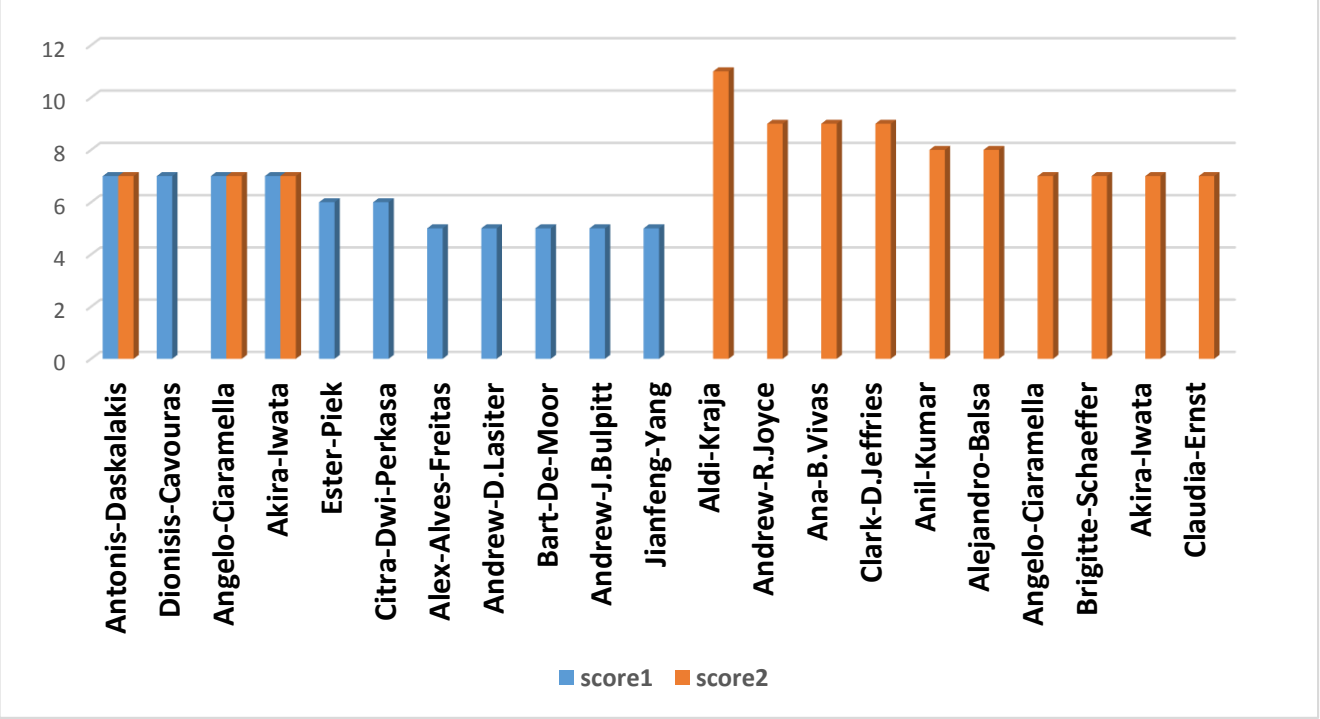


Figure 3.8 Comparison between 1st and 2nd execution results.

Score1 in bleu represents the scores obtained during the first execution (first launch of *PaToDet*) and score2 in red is related to the second execution (second launch of *PaToDet*).

The results of our experiment are promising, but depend upon the selected dataset and the generated tokens related keywords. At least the proposed approach enables the detection of the exact content cascade as in [36].

III.6 Future work

As future works, we intend to refine the process of extending the context of the frequent content-tokens in order to obtain more relevant related keywords, and develop more the experimental works. We also propose to use some advanced algorithm techniques in order to gain time by reducing the execution time, for example use memory mapping and parallel techniques.

III.7 Conclusion

To sum all the work up, we have proposed an approach to detected influencers based on the previous works using information cascade. We believe that the proposed approach could lead to the detection of exact content-tokens propagation as well as their related tokens. The results of our experiments are promising, but depend upon the selected dataset and the generated tokens related keywords. At least the proposed approach enables the detection of the exact

content cascade as in [35]. More efforts should be done to generate more relevant content-tokens related keywords.

As future works we intend to refine the process of extending the context of the frequent content-tokens in order to obtain more relevant related keywords, and develop more the experimental works.

Conclusion

Influencers detection has attracted the attention of researchers, politicians, companies, and governmental institution, It is a very active line of research due to its several real-world applications. Content-based networks analysis was not extensively investigated, more work could be done in this direction.

In this work, we have proposed an approach to detect influencers in social networks inspired from a previous work based on information cascade. We believe that the proposed approach could lead to the detection of exact content-tokens propagation as well as their related tokens. At least the proposed approach enables the detection of the exact content cascade. More efforts should be done to include semantics in studying information dissemination in networks. The experiments presented in this manuscript were conducted on DBLP that was assimilated to a social network. This work could be generalized to social networks, we believe that the idea of using Google search API and Alchemy APIs would give weak results when applied to natural language, a more accurate solution is to use in this case WordNet API which could lead to more efficient results.

References

[1] Linton C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40(1):35 41, 1977.

[2] Duncan J. Watts and Steven H.Strogatz. Collective dynamics of small world networks. Nature, 393:440 442, 1998.

[3] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. Sociometry, 32(4):425 443, 1969.

[4] https://sysomos.com/inside-twitter/twitter-friendship-data

[5] Mark Granovetter. The strength of weak ties. American Journal of Sociology, 78:1360–1380, 1973.

[6] Newman, M. E. J. (2001). Clustering and Preferential attachment in growing networks. PhysicalReviewLettersE,64 (025102).

[7] P.W.Hollandand, S. Leinhardt. Transitivity in structural models of small groups. Small Group Research, 2:107124, 1971.

[8] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. Nature, pages 440–442, Jun 1998.

[9] Adrien Guille, Hakim Hacid, Cécile Favre, Djamel A. Zighed Information Diffusion in Online Social Networks: A Survey. SIGMOD Record, June 2013 (Vol. 42, No. 2)

[10] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu , social media mining, 2014. April 20, 2014.

[11] D. Easley and J. Kleinberg. Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press, 2010. Draft version: June 10, 2010.

[12] E. M. Jin, M. Girvan, and M. E. J. Newman. Structure of growing social networks. Phys. Rev. E, 64:046132, 2001.

[13] P. Reddy et al. A graph based approach to extract a neighborhood customer community for collaborative filtering. In DNIS, 2002.

[14] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In SIGIR, 2006.

[15] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In WWW, 2012.

[16] Chi Wang, Jie Tang, Jimeng Sun, and Jiawei Han. Dynamic social influence analysis through time-dependent factor graphs. In Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on, pages 239–246. IEEE, 2011.

[17] Karthik Subbian et al Social capital: the power of influencers in networks. In Proceedings of the international conference on Autonomous agents and multi-agent systems, pages 1243–1244, 2013.

[18] Charu Aggarwal and Karthik Subbian. Evolutionary network analysis: A survey. ACM Computing Surveys (CSUR), 47(1):10, 2014.

[19] C. Kiss and M. Bichler, "Identification of influencers – Measuring influence in customer networks," Decision Support Systems, vol. 46, pp. 233–253, Dec. 2008.

[20] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million-follower fallacy. In ICWSM-10: Proceedings of international AAAI Conference on Weblogs and Social, 2010.

[21] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, asocial network or a news media? In Proceedings of the 19thinternational conference on World Wide Web, WWW '10,pages 591{600, New York, NY, USA, 2010. ACM.

[22] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In KDD, pages 1029–1038. ACM, 2010.

[23] Prem Melville , Karthik Subbian , C Perlich ,R Lawrence, and E Meliksetian : A predictive perspective on measures of influence in networks. In Proceedings of the Workshop on Information in Networks, 2010.

[24] Karthik Subbian and Prem Melville. Supervised rank aggregation for predicting influencers in twitter. In Social Com, pages 661–665. IEEE, 2011.

[25] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In KDD '10, pages 1019–1028, 2010.

[26] J. Makkonen, H. Ahonen-Myka, and M. Salmen kivi. Simple semantics in topic detection and tracking. Inf. Retr., 7(3-4):347–368, Sept. 2004.

[27] F. Wang, H. Wang, and K. Xu. Diffusive logistic model towards predicting information diffusion in online social networks. In ICDCS '12 Workshops, pages 133–139, 2012.

[28] J.Weng, E.-p.Lim, and J.Jiang," Twitter rank: Finding Topic-Sensitive Influential Twitterers Twitter Rank : Finding Topic-sensitive Influential Twitterers," pp. 261–270, 2010.

[29] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In KDD '09, pages 497–506, 2009.

[30] D. A. Shamma, L. Kennedy, and E. F. Churchill. Peaks and persistence: modeling the shape of microblog conversations. In CSCW '11, pages 355–358, (short paper) 2011.

[31] Philip E. Brown, Junlan Feng Measuring User Influence on Twitter Using Modified K-Shell Decomposition. 2011, Association for the Advancement of Artificial Intelligence.

[32] J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," in Proceedings of the ninth annual ACM-SIAM symposium on discrete algorithms, no. May 1997, pp. 668–677,1998.

[33] A. Pal and S. Counts. Identifying topical authorities in microblogs. In WSDM '11, pages 45–54, 2011.

[34] Greg Ver Steeg & Aram Galstyan : information theoretic measures of influence based on content dynamics.WSDM'13 Proceedings of the sixth ACM international conference on web search and data mining, page 3-12 ,2013.

[35] karthik subbian , Jaideep Srivastava,Arindam Banerjee, Scalable Analysis of Information Flow in Networks, page 19-26, 2014.

[36] Karthik Subbian, Charu Aggarwal, and Jaideep Srivatsava. Mining influencers using information flows in social streams. ACM Trans. Knowl. Discov. Data 10, 3,Article 26 (January 2016), 28 pages. DOI: http://dx.doi.org/10.1145/2815625.

[37] Singhal Ayush. (2014). Leveraging open source web resources to improve retrieval of low text content items. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/168312.

[38] https://developers.google.com/apis-explorer/#p/.

[39] http://www.alchemyapi.com/.

[40] http://dblp.uni-trier.de/xml/.