Social Network Analysis and Information Propagation: A Case Study Using Flickr and YouTube Networks

Samir Akrouf, Laifa Meriem, Belayadi Yahia, and Mouhoub Nasser Eddine, *Member, IACSIT*

***Abstract*—Social media and Social Network Analysis (SNA) acquired a huge popularity and represent one of the most important social and computer science phenomena of recent years. One of the most studied problems in this research area is influence and information propagation. The aim of this paper is to analyze the information diffusion process and predict the influence (represented by the rate of infected nodes at the end of the diffusion process) of an initial set of nodes in two networks: Flickr user's contacts and YouTube videos users commenting these videos. These networks are dissimilar in their structure (size, type, diameter, density, components), and the type of the relationships (explicit relationship represented by the contacts links, and implicit relationship created by commenting on videos), they are extracted using NodeXL tool. Three models are used for modeling the dissemination process: Linear Threshold Model (LTM), Independent Cascade Model (ICM) and an extension of this last called Weighted Cascade Model (WCM). Networks metrics and visualization were manipulated by NodeXL as well. Experiments results show that the structure of the network affect the diffusion process directly. Unlike results given in the blog world networks, the information can spread farther through explicit connections than through implicit relations.**

***Index Terms*—Information diffusion, influence, social media, social network analysis.**

I. Introduction

Human social relationships were bounded according to time and space, but the evolution of information and communication technologies tools allowed people to inexpensively and reliably share information anytime and anywhere through social media (YouTube, Flickr, Twitter, Facebook, blogs, emails, etc). These tools are helpful recourses of information, opinions and behaviors regarding different areas of interest. Studying and measuring these social media have attracted considerable interest of many researchers in various domains and led them to create a new field called Social Network Analysis (SNA). SNA methods have been applied to a wide range of areas like business, healthcare, academia, politics and terrorism [4], [5], [8], [9], [10], [12], [14].

In our daily life, there are innumerable situations in which we are influenced in our decision making by what others around us are doing. Simple examples of influence are when academic researchers choose to work on a topic that is currently ″hot″, or when we listen to the same music that our friends listen to. To study this kind of decision making, Banerjee [1] has developed a concept in which a person makes decisions based on what other people do because their decisions may reflect information that they have and he or she does not. This concept is called ″herding″ or ″information cascades″. Therefore, analyzing the flow of information on social media and predicting users' influence in a network became so important to make various kinds of advantages and decisions. In [2]-[3], the marketing strategies were enhanced with a word-of-mouth approach using probabilistic models of interactions to choose the best viral marketing plan. Some other researchers focused on information diffusion in certain special cases. Given an example, the study of Sadikov et.al [6], where they addressed the problem of missing data in information cascades, and evaluated their methodology using information propagation cascades in Twitter network, by involving a K-tree model to estimate properties of the cascade of information, such as size and depth. Moreover, a study on the Blog worlds proposed a special model of information diffusion based on explicit and implicit links [7]. Explicit links are the relations formed between blogs directly to obtain information or to maintain a relationship, whereas when information is diffused between blogs not through an explicit relationship, it is called implicit link. Other researchers have studied and modeled social media epidemics (like viruses and rumors) especially on Twitter [11]-[13]. Several mathematical and physical based diffusion models have been suggested to formally the spread of information in a network [15]-[16]. The literature offers four basic approaches to modeling influence propagation in social networks: cascade models, threshold models, epidemic models and game theory models [17]-[18].

In this paper, we focus on analyzing the information propagation process for anticipating the capability of nodes in spreading the information throughout the network. We also aim to understand how the structure of the network and the type of its relationships can influence the propagation process. Likewise, this analysis is done on two different networks: an explicit network created from Flickr user's contacts, and an implicit network created from users' comments on YouTube videos. The discussion of networks treats them as static structures: we take a snapshot of the nodes and edges at a particular moment in time and then analyze their structure and the diffusion of information process. These networks were extracted using NodeXL tool.

The rest of the paper is organized as follows: the second section describes the implemented diffusion models used to predict the influence in the networks. The third section is devoted to the tools and the used datasets. We also compare the structure of the used networks in this section. The results of the experiments are given in this section as well. Section 4 discusses the obtained results and finally a conclusion is given in section 5.

Manuscript received September 4, 2012; revised November 18, 2012.

Samir Akrouf, Laifa Meriem, Belayadi Yahia, and Mouhoub Nasser Eddine are all with the Universite de Bordj Bou Arreridj Algeria (e-mail: samir.akrouf@gmail.com).

II. THE MODELS

Given a network G = (V, E) where V is the set of vertices, and E the set of existing edges in the network. A vertex $v \in V$ is said to be active if the information has reached the vertex and was accepted by it. If the information didn't reach the vertex or the vertex rejected it, then the vertex is said to be inactive. Each inactive vertex tends to become active, and it can switch from inactive to active, but it does not switch in the other direction. There must be an initial set of vertices activated to start the diffusion process targeted for initial activation. They are called 'initial adopters' of the information. The influence of this initial set of vertices is the expected number of active vertices in the end of the diffusion process. Consequently, the cascading process will appear as follows: given an initial set of active vertices; while time spreads out, more of an inactive vertex v's neighbors become active which may cause this vertex to become active at some point. Then v may in turn trigger other vertices to which it is connected to adopt the same decision or action. We have chosen two basic diffusion models: the Linear Threshold Model (LTM) and the Independent Cascade Model (ICM).

III. LINEAR THRESHOLD MODEL

One basic approach to model information dissemination in networks is based on the use of node-specific thresholds. Granovetter was among the first to propose threshold models in sociology [19]. We first implement a generalization of the LTM proposed by Kempe et.al [20]. In this model, a vertex v is influenced by each neighbor w according to a weight b_{vw} where $b_{vw} \leq 1$. Each vertex v chooses a threshold T_v uniformly and randomly from the interval [0, 1]. This threshold is defined as "the weighted function of v's neighbors that must become active before v becomes active". The random choice of the thresholds T, at each time the process of diffusion runs; it fills the lack of information about the network and the relations between its actors. As explained before, the diffusion process unfolds in discrete steps: in step s: all vertices that were active in step (s-1) remain active, and new vertex v is activated if the total weight of its active neighbors is at leastT_v: $\sum_{w \in N_v} b_{vw} \geq T_v$. The information stops propagating when there are no more inactive vertices that can become active.

Given two vertices u, v; if c_{uv} is the number of parallel edges between them (the multiplicity of the edge e_{uv} that is treated as weights), and d_u, d_v are the overall degrees of u and v, and then the weights b_{uv} and b_{vu} are computed as follows:

$$b_{uv} = \frac{c_{uv}}{d_v} \quad, and \ b_{vu} = \frac{c_{vu}}{d_u}$$

In other words, it is the ratio of an edge to permit the spread of information from a node to its neighbors. Giving a simple example from real world; a new book comes on the libraries and several friends buy it. As much of your friends read this book, they will eventually convince you to read it as well. This is how LTM works.

IV. INDEPENDENT CASCADE MODEL AND WEIGHTED CASCADE

The second model is based on the work in interacting particle systems which describe the behavior of systems by probability theory [21]. ICM starts with an initial set of active vertices A_0, and the process unfolds in discrete steps according to the following randomized rule. When vertex v first becomes active in step s, it is given a single chance to activate each of its inactive neighbors w with a probability P_{vw} to succeed. If v succeeds to activate w in step s+1, then w will try to activate its inactive neighbors too. Otherwise, v cannot make any further attempts to activate w in subsequent rounds. In case that w has many new active neighbors, they attempt to activate w in an arbitrary order. The probability P_{vw} is an independent parameter of the system [20]. At first, each edge in the network was assigned with a uniform probability. We chose the success probability p to be 50% that gives equal chances to a vertex to be successfully activated or not. The diffusion process ends when there are no more inactive vertices that can be switched to become active at a step s. If we use the books example from the ICM point of view, then we can formulate it as follows: you were just convinced by the new book and you bought it. Then, you will talk about it to your friends and try to convince them to buy it too, but you will have only one chance to convince them and you can never try again.

A special case of ICM proposed also in [20] is implemented too, where each edge from a vertex u to v has a success probability $p = \frac{1}{d_v}$ for activating v, where d_v is the overall degree of the vertex v. This model is called the "Weighted Cascade".

V. EXPERIMENTS

A. Tools and Datasets

The information propagation process was analyzed on two different networks: an explicit network from Flickr social media, and an implicit network from the social service YouTube. Both networks were unimodal and extracted using the free and open tool NodeXL[22]. NodeXL allows visualizing and analyzing networks graphs, and computing the network metrics as well [23].

Flickr allows users to upload and share digital photos, and recently videos. Flickr is a reflection of the society through images and the community that has emerged with it and around it. Therefore, it is interesting to be analyzed in many purposes like: personal network analysis, community sphere, E-commerce, Service and Infrastructure, Geo-Tagged Applications. Interactions on Flickr can be either *explicit* (e.g. belonging to the same group, or adding each other as contacts), or *implicit* (e.g. commenting on each other's photos, assigning the same tags to a photo). In our experiments we used an explicit contacts network from a public profile that gave 2534 nodes representing Flickr users (contacts of the contact), and 2650 links representing the contact relationship with a weight equal to 1 (a user can add another contact only once). The obtained network is directed and egocentric.

YouTube is one of the most popular social media dedicated to share videos. Video content is used for many purposes, from conveying knowledge and disseminating information to self-promotion to documenting world affairs. Several types of networks can also be extracted from YouTube Social media. We used an implicit network created from users' comments on videos (only videos that related to education and science). Nodes in this network represent videos and a link between a pair of videos is created only if a user commented on both videos. Since a user can comment more than once on videos, we needed to merge all duplicated edges, and the multiplicity of links is treated as weights. The final given network had 836 nodes and 3885 links between them. It is a directed non-egocentric and implicit network.

TABLE I: STRUCTURE CHARACTERISTICS OF THE USED NETWORKS

	Flickr	YouTube
Graph type	Directed	Directed
Egocentric network	Yes	No
Type of relationship	Explicit	Implicit
Number of vertices	2534	836
Number of edges	2650	3885
Graph density	0.0004	0.0055
Connected component	1	41
Maximum number of vertices in connected component	2534	671
Maximum number of edges in connected component	2650	3707
Diameter	4	16
Average distance	3.775738	5.69307

TABLE II: METRICS' STATISTICS

	Flickr	YouTube
Min Overall Degree	1	1
Max Overall Degree	301	44
Average Overall Degree	2.087	9294
Min Out-Degree	0	0
Max Out-Degree	300	33
Average Out-Degree	1041	4647
Min In-Degree	1	0
Max In-Degree	6	38
Average In-Degree	1046	4647
Min Betweenness	0	0
Max Betweenness	369816.207	183326.076
Average Betweenness	7034.719	2545.385

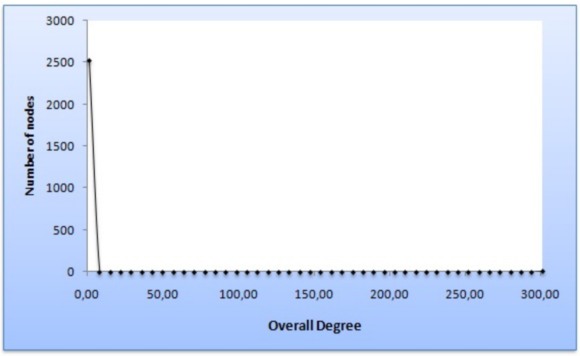

Fig. 1. Flickr contacts network overall degree distribution.

Table I and Table II summarize the structure characteristics and the computed centrality metrics for both networks. The used networks are different, in their size, types, and even relationships. The overall degree distribution in the explicit contacts network from Flickr shows that most nodes (98%) have a low overall degree (see Fig.1) due to the type of the network, where we found only the ego and its directed contacts with significant degree values. Indeed, the comments network from YouTube gave a different plot in Fig.2. The average in-degree and out-degree is *4.647*. This is explained by the density of the network and its large diameter. One should also notice that the maximum in-degree and maximum out-degree are 38 and 33 respectively, which was expected in a non-egocentric network.

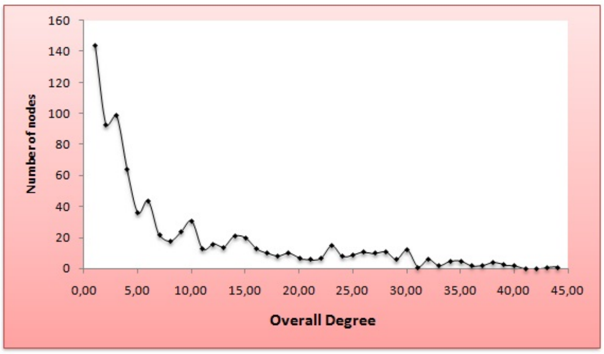

Fig. 2. YouTube comments overall degree distribution.

B. *Initial Active Set*

Another important step in the implementation of the chosen models is "choosing the initial active set" or the initial adopters. The problem is posed as follows: "if you want to trigger a large cascade of adoptions of a new product or innovation, you need first to convince a subset of individuals to do so. In this case, which set of individuals (nodes) should you target initially? ". As we mentioned before, we focus in this work only on the structure of the network. So, the initial active set is chosen based on the structural characteristics of nodes.

The first initial active set is chosen based on the high degree centrality metric which is a count of total number of connections linked to a node. In real world networks, the degree centrality measures how many people a person can reach directly in the network. Since the used networks in the experiments are all directed, the degree is not computed but the in-degree and out-degree for each node. In-degree represents how many links those point inward at a node, while the out-degree is the number of links those point outward to other nodes, and their sum is called overall degree. We choose the initial adopters based on the high out-degree and the high overall degree values.

The second choice is made based on the high betweenness centrality values. Betweenness is another important metric that ranks the importance of a node according to its position in the network. It is the number of shortest paths from all nodes to all others that pass through that node. The idea is that actors who are "between" other actors, and on whom other actors must depend to reach others are more important in a social network. Calculating the betweenness centrality for all nodes can be very costly especially in large networks. Therefore, many studies have focused on developing new algorithms to reduce the time and space cost. The used algorithm implemented on NodeXL is developed by U. Brandes [24], which requires $O(n + m)$ space and runs in $O(nm + n^2 \log n)$ time on weighted graphs. Note that n is the number of nodes in the network and m is the number of links.

Finally, as a basic and intuitive choice, the same algorithms are tested by choosing the initial active set randomly with no specific condition. This choice will permit to compare and prove the important role of choosing the initial adopters for spreading the information widely in a social media.

The size of initial active set depends on the size of network, its type, and the domain of the experiments. For example: in marketing; targeting a large size of initial adopters set can be costly since one has to pay each initial adopter. In the other side, spreading a rumor in Twitter, for example, doesn't require any cost, then targeting a large initial adopters set will be better. In our experiments, the size of initial active nodes is chosen to be '20'.

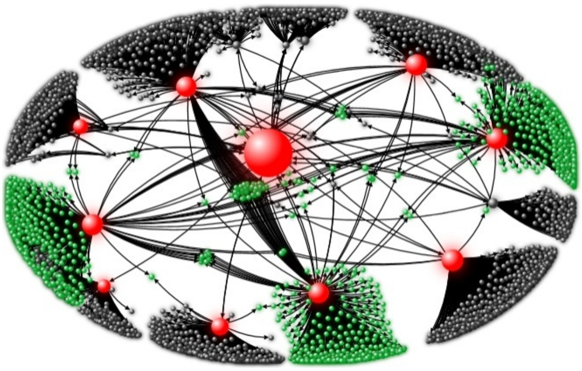

Fig. 3. An instance after a diffusion run using the ICM on the Flickr explicit network.

C. Results

The selected diffusion models were carried out (i.e., implemented) on both datasets: the Flickr contacts network as an explicit network, and the YouTube comments network as an implicit network. The results are presented in this section from both datasets. For each algorithm, the propagation process was run 1000 times for every initial set $A_0 \in \{5, 10, 15, 20\}$, then the average of the size of active nodes after each run was computed. This average is considered to be the influence of the initial set. Fig.3 illustrates an instance of diffusion process applied on the Flickr explicit network using the independent cascade model by choosing 10 initial active nodes (in red) based on the high betweenness centrality. NodeXL enables visualizing the nodes based on their metrics. By the end of the process, the initial adopters have influenced about 35% of the nodes (infectious nodes are green). Figures 4, 5 and 6 show the performance of the algorithms in the LTM, ICM and WCM respectively on the Flickr contacts network. At the first sight of these plots, we noticed, for each diffusion model, the results converge to be proximate and close when the initial active set is chosen based on overall degree, out-degree or betweenness centrality. We also noticed that the influence ratio of random initial adopters is very low compared to other cases (as expected, choosing random initial adopters is not a good choice).

Another observation is that the curves become steady when the initial active set contains more than 15 nodes. There is a natural explanation to this observation. The first 15 targeted nodes (with high overall degree, out-degree or even betweenness centrality) influence a large fraction of the network. However, the additional 5 nodes reach only a small supplemental fragment of the network since the nodes with higher values are already active. For the LTM, the influence of nodes with high overall degree, out-degree and betweenness centrality ended by giving the same high influence ratio (96%), while the random initial active set ended by infecting only about 25 nodes. WCM surpasses, as well, by more than 90% for different initial active set except the random choice.

To investigate the reason why both models gave such results on this network, we have been observing the network after each execution. The justification is that both models depend on the overall degree of a node to activate it or not. In the LTM, the node becomes active if the sum of the weight of the incoming edges from its active neighbors divided by its overall degree is equal or greater than its threshold. On the other hand, in WCM, a node can influence its neighbor with a success probability $p = 1\backslash d_v$ (d_v is the overall degree of the target node). As the overall degree distribution plot shows (Fig.1), most nodes in the network have a very low overall degree. This causes the activation function in the LTM and the success probability in the WCM to be of high values. Therefore, the influence rate was significant from both models.

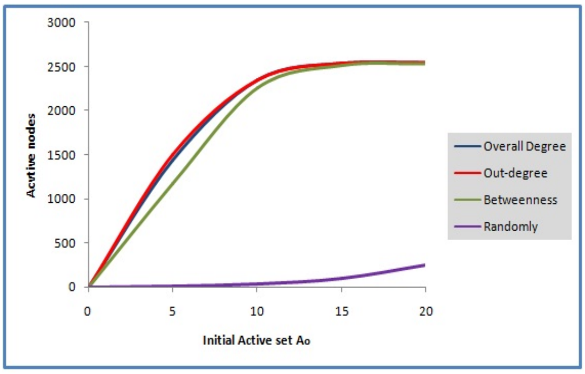

Fig. 4. Results for LTM (Flickr).

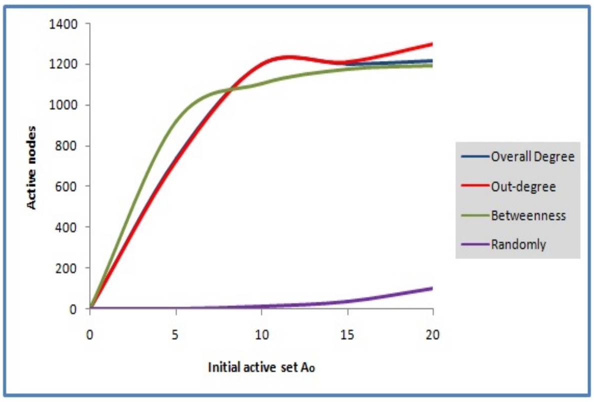

Fig. 5. Results for ICM (Flickr).

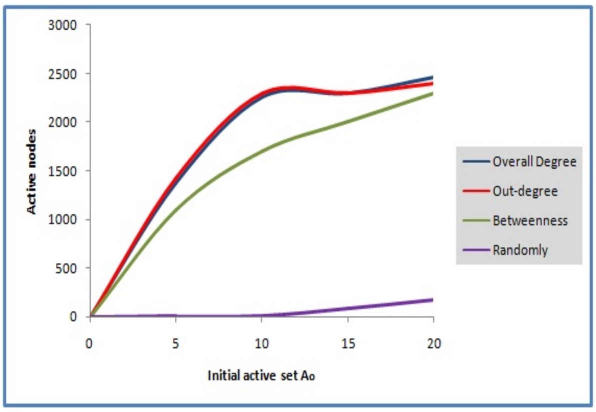

Fig. 6. Results for WCM (Flickr).

Fig. 5 shows the results for the ICM (with a success probability $p = 0.5$) that look slightly different than both other models. When the initial active nodes are less than 8, the heuristic of choosing nodes with high betweenness centrality values seems to perform better than other heuristics. But you can notice that by the end of the experiments, the out-degree heuristic outperforms all other choices. One can also notice that the scale is smaller. The influence ratio from this model does not exceed 51%. The first explanation that came to mind is the strength and weakness of the links (ties) between nodes. Stronger ties represent greater frequency of interaction in our experiments. Since the Flickr network (i.e., the corresponding Flickr dataset) was taken as a snapshot at a specific moment, the frequency of interactions was fixed to 1 as the user can add another contact only once. To prove this substantiation we needed to compare it with the YouTube comments network where the frequency of interactions is not fixed to one value.

Fig. 7, 8 and 9 show the diffusion process results from the experiments on the implicit network. The initial surprising remark about these results is the general low influence rate compared to the Flickr network results. This was not expected since the network density and diameter are greater. To understand these results, we have observed the network after many tests. Fig.10 illustrates the diffusion process resulting from the LTM by targeting 10 initial active nodes (red) based on their high out-degree values. The nodes are visualized by the in-degree values. After filtering nodes by displaying only vertices with in-degree greater than 4, the infected nodes (pink) seem to be with high in-degree values. For more analysis, we filtered nodes based on their thresholds, and found out that most influenced nodes have thresholds less than 0.3 (Fig.11).

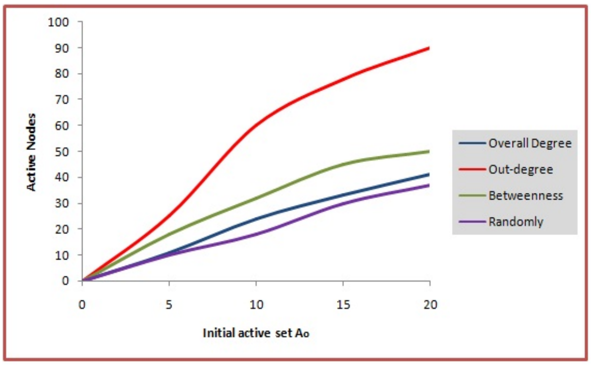

Fig. 7. Results for LTM (YouTube).

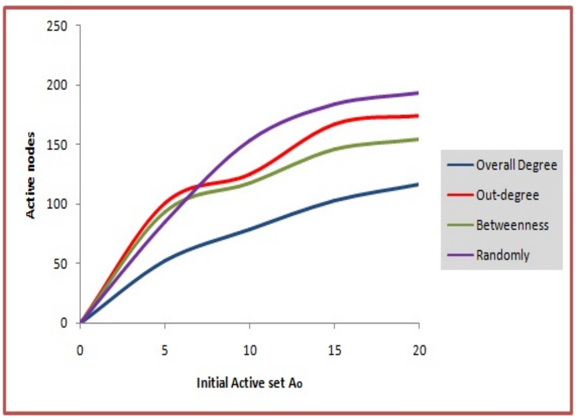

Fig. 8. Results for ICM (YouTube).

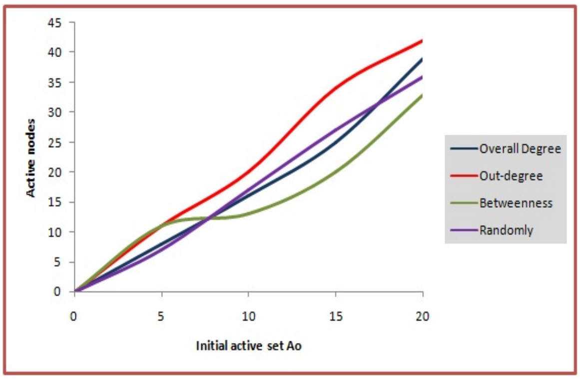

Fig. 9. Results for WCM (YouTube).

Starting with results for the LTM in Fig.7; this model outperformed when the initial adopters with high out-degree values were chosen to launch the propagation process and ended by influencing only 90 nodes in the network. Betweenness centrality, overall degree and random initial active set tests gave close results with very low influence ratio (0.06%, 0.05% and 0.04% respectively). This is due to the fact that most nodes have a high overall degree, which causes the activation function to be of low values especially when most neighbors of a vertex are not active. Another unexpected result is illustrated in Fig.8, where one can notice that the random active set curve surpasses all other heuristics for the ICM with a success probability always p = 0.5. A snapshot was taken to understand this case in Fig.12. The initial active nodes not only belong to the giant component but to another small component which was not possible with the other heuristics. Despite the fact that the initial adopters do not have high out-degree values (the size of nodes is set to be their out-degree values), they infected about 279 nodes (more than 23% of the network), while the overall degree heuristic infected only 14% of the whole network. It is interesting to notice that the rate of diffusion on the giant component only is more than 29% which is higher than the influence ratio on the whole network.

For the last model: WCM (in Fig.9), betweenness centrality based heuristic seemed to perform better when only 5 nodes are targeted to start the diffusion process. When the initial active set has more than 5 adopters, the out-degree heuristic outperformed but with a very low influence ratio compared to the previous models. The reason why this model did not carry out as well as the ICM is that the overall degree of nodes in this network is of an average 9. This means the average success probability of a node to be active is around 1/9 which is weak enough.

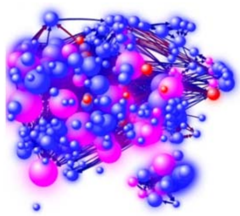

Fig. 10. Filtering based on the in-degree values.

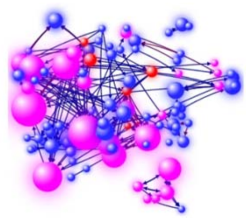

Fig. 11. Filtering based on the threshold values.

VI. Discussion

The higher influence prediction rates from the experiments were given by the LTM and the WCM in the case of explicit egocentric networks because of the low degree of nodes in scale-free networks, while the ICM performed better on implicit networks with stronger ties since it is based on the interactions between nodes but not the overall degree values. In this section, we discuss the obtained results of the information diffusion process in both networks: Flickr contacts network, and YouTube comments network.

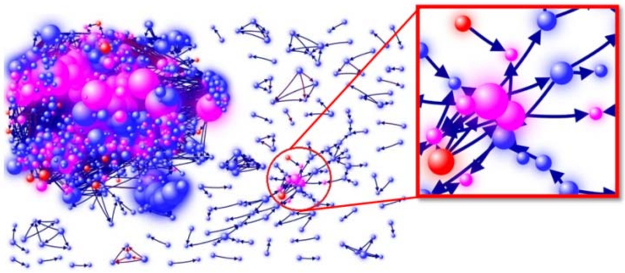

Fig. 12. An instance after a diffusion run using the LTM on the YouTube implicit network where the initial adopters are chosen randomly.

The network structure is the main factor that impacts the process of information dissemination on networks. As experiments on YouTube comments network demonstrated, the flow of information can be restrained in few components of a network. This recalls the importance of local bridges in networks for spreading the information.

Another reason for the low influence rates in the YouTube comments network (in addition to the isolated components) is its large diameter compared to the Flickr contacts network. The Diameter is an important metric of network performance that can also affect the flow of information in social media. Our experiments have proven that potential capacity of information diffusion in Flickr contacts network is quite higher as the average distance and the network diameter stand at low value, which was already proven in blogosphere [25]. Some researchers worked on minimizing the network diameter to allow a wide spread of information [26]. From the YouTube comments network experiments, we noticed that the influence ratios were higher in the giant component compared to the ratios in the whole network, which may be explained by the fact that information can disperse larger in dense networks but yet the results need more experiments in denser networks to be proven.

Another main observation is that the information can spread wider in egocentric and explicit network (in our case it is a Flickr contact network). A possible reason to this large information dissemination is the principle that people who like each other's pictures on Flickr tend to add each other as contacts to become friends, and people who are friends tend to like each other's pictures and share it with their friends.

YouTube interactions hold in several kinds of relationships which construct different types of networks. Many researchers have analyzed information diffusion process in YouTube explicit networks that contain friendship or subscription relationships or both [27], [28]. Our experiments proved that commenting is also another important relationship that may spread information and influence in social media. However, when more than 85% of all information diffusion in a blog world happens through non-explicit relationships [7], only about 18% of information can spread through comments on YouTube videos, due to the fact that people tend to share the video they find interesting or to like it rather than commenting. We should also keep in mind that users' comments are not always helpful for spreading information. This demand farther work to take into account the type of the comment (with the information or against it).

Finally, the experiments evinced that triggering a diffusion process in social media based only on the structure properties of nodes can be tricky (as seen with the random heuristics on the YouTube comments network) especially when the centrality metrics distribution of the network's nodes does not follow a power law distribution. As seen before, centrality metrics heuristics performed well enough when the network structure is not complex, but the results were completely unexpected when the network structure was more complex. Thus one should take into account all the network characteristics before choosing the initial adopters.

VII. Conclusion

In this work, we analyzed information diffusion process and the influence of a set of nodes in two different networks: an egocentric contacts network created by explicit relationships from Flickr social service, and an implicit videos network created by commenting relationship from YouTube service. These networks were extracted and visualized using NodeXL tool. Information propagation was modeled using the Linear Threshold Model (LTM), the Independent Cascade Model (ICM), and an extension from this latter called Weighted Cascade Model (WCM). To trigger the diffusion process, an initial set that contains active nodes had to be chosen. Selecting these active nodes was based on the structural metrics of nodes, where nodes with high overall degree, out degree and between centrality values were chosen. Another selection was random to compare its influence rates with the other well structural criteria. The experiments have shown that choosing the initial adopters based only on the centrality metrics can be tricky.

However, our analysis of propagation process was based on the structure of the networks and the type of their relationships. We first noticed that choosing the diffusion model is relevant to the network and its type of relationships, where LTM and WCM were better to be used on egocentric explicit networks to get better influence predictions, while ICM performed better on implicit networks with stronger ties since it is based on the interactions between nodes.

Indeed, we also confirmed that the low value the average distance and the network diameter allow information to spread larger in social networks. Unlike information propagation in the blog world network, our experiments have established that commenting implicit relationship in YouTube social service limits the spread of information.

References

[1] A. V. Banerjee, "A simple model of herd behavior," *The Quarterly Journal of Economics*, vol. 107, no. 3, pp. 797-817, August, 1992.

[2] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining ACM New York*, NY, 2001.

[3] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proc. of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining ACM New York*, NY, USA ©2002.

[4] V. Krebs, "Working in the connected world: Social capital — The KillerApp for HR in the 21st Century," *IHRIM Journal*, June 2000.
[5] V. Krebs, "Managing the 21st Century Organization," *IHRIM Journal*, vol. 11, no. 4, 2007.
[6] E. Sadikov, M. Medina, J. Leskovec, and H. G. Molina, "Correcting for missing data in information cascades," *WSDM'11*, Hong Kong, China, February 9–12, 2011.
[7] Y. S. Kwon, S.W. Kim, S. J. Park, S. H. Lim, and J. B. Lee, "The information diffusion model in the blog world," *The 3rd SNA-KDD Workshop '09 (SNA-KDD'09)*, Paris, France, June 28, 2009.
[8] A. L. Barabasi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica*, no. 3-4, pp. 590-614, 2002.
[9] M. Andre, K. Ijaz, J. D. Tillinghast, V. E. Krebs, L. A. Diem, B. Metchock, T. Crisp, and P. D. McElroy, "Transmission network analysis to complement routine tuberculosis contact investigations," *American Journal of Public Health*, vol. 96, no. 11, November 2006.
[10] K. Jafa, P. McElroy, L. Fitzpatrick, C. B. Borkowf, R. MacGowan, A. Margolis, K. Robbins, A. S. Youngpairoj, D. Stratford, A. Greenberg, J. Taussig, R. L. Shouse, M. LaMarre, E. M. Lemal, W. Heneine, and P. S. Sullivan, "HIV transmission in a state prison system 1988–2005," PLoS ONE, vol. 4, no. 5, May 2009.
[11] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Z. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1589–1599, Edinburgh, Scotland, UK, July 27–31, 2011.
[12] S. D. McClurg, M. L. Wade, and M. V. W. Phillips, "Gender, social networks, and voting behavior, Political networks paper archive," *Working Papers*, pp. 64, 2012.
[13] P. V. Fellman, R. Wright, "Modeling terrorist networks - complex systems at the mid-range," *Complexity*, *Ethics and Creativity Conference*, *LES*, 2003.
[14] M. Broecheler, P. Shakarian, and V. S. Subrahmanian, "A scalable framework for modeling competitive diffusion in social networks," *Social Computing (SocialCom)*, *IEEE Second International*, 2010.
[15] H. Ma, H. X. Yang, M. R. Lyu, and I. King, "Mining social networks using heat diffusion processes for marketing candidates selection," *CIKM'08, Napa Valley, California, USA*, October 26–30, 2008.
[16] D. Easley and J. Kleinberg, "Networks, crowds, and markets: Reasoning about a highly connected world," Cambridge University Press, 2010.
[17] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, pp. 167-256, 2003.
[18] M. Granovetter, "Threshold models of collective behavior," *The American Journal of Sociology*, vol. 83, no. 6, pp. 1420-1443, May, 1978.
[19] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," *KDD '03 Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining ACM New York*, NY, 2003
[20] T. M. Liggett, "Interacting particle systems- An introduction," *School and conference on probability theory*, Trieste 13-31, May, 2002.
[21] NodeXL Graph Gallery. [Online]. Available: http://nodexl.codeplex.com/
[22] D. L. H. B. Shneiderman and M. A. Smith, "analyzing social media networks with nodexl," *Insights from a Connected World*, Elsevier, 2011.
[23] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163-177, 2001.
[24] Z. Babutsidze, T. Lomitashvili and K. Turmanidze, "The structure of georgian blogosphere and implications for information diffusion," August 5, 2011.
[25] E. D. Demaine and M. Zadimoghaddam, "Minimizing the diameter of a network using shortcut edges," *Lecture Notes in Computer Science*, vol. 6139, Algorithm Theory - SWAT 2010, pp. 420-431, 2010.
[26] J. C. Paolillo, "Structure and network in the YouTube core," in *Proc. of the 41st Annual Hawaii International Conference on System Sciences*, pp. 156, IEEE Computer Society, Washington DC, USA ©2008.
[27] J. I. Biel, "Please, subscribe to me! Analysing the structure and dynamics of the YouTube network," *Idiap Research Institute, Ecole Polytechnique Federale, Lausanne*, 2009.

Samir Akrouf was born in Bordj Bou Arreridj, Algeria in 1960. He received his Engineer degree from Constantine University, Algeria in 1984. He received his Master's degree from University of Minnesota, USA in 1988. He received his Phd degree from University of Setif Algeria. Currently; he is an assistant professor at the Computer department of Bordj Bou Arréridj University, Algeria. He is an IACSIT member and is a member of LMSE laboratory (a research laboratory in Bordj Bou Arreridj University). He is also the dean of Mathematics and Computer Science Facultyy of Bordj Bou Arreridj University. His main research interests are focused on Biometric Identification, Computer Vision and Computer Networks and social network analysis.

Meriem Laifa was born in Setif , Algeria in 1988. She received her bachelor degree from Setif University in 2010. She received her Master's degree from Bordj Bou Arreridj University in July 2012. Now she is a Phd student.

Yahia Belayadi was born in Bordj Bou Arreridj, Algeria in 1961. He received his Engineer degree from Setif University Algeria in 1987. He received his magister from Setif University Algeria in 1991. Currently he is an assistant professor at the Computer department of Bordj Bou Arreridj University, Algeria. He also is the director of University Center of Continuous Education in Bordj Bou Arreridj.

Nasser Eddine Mouhoub was born in Bordj Bou Arreridj, Algeria in 1962. He received his Engineer degree from Setif University Algeria in 1988. He received his magister from Setif University Algeria in 2003. He received his Phd degree from Setif University Algeria in 2011. Currently; he is an associate professor at the Computer department of Bordj Bou Arreridj University.