Influential Actors Detection Using Attractiveness Model in Social Media Networks

Ziyaad Qasem, Marc Jansen, Tobias Hecking and H.Ulrich Hoppe

Abstract Detection of influential actors in social media such as Twitter or Facebook can play a major role in improving the marketing efficiency, gathering opinions on particular topics, predicting the trends, etc. The current study aspires to extend our formal defined T measure to present a new measure aiming to recognize the actors influence by the strength of attracting new attractors into a networked community. Therefore, we propose a model of an actor influence based on the attractiveness of the actor in relation to the number of other attractors with whom he/she has established connections over time. Using an empirically collected social network for the underlying graph, we have applied the above-mentioned measure of influence in order to determine optimal seeds in a simulation of influence maximization.

1 Introduction

With the wide spread of social media networks nowadays, it has become possible to acquire insights into and knowledge about a wide variety of more or less numerous communities interacting through the Internet. Moreover, applying analytic approaches to social media data can provide better-informed decision-making processes in various fields like marketing, politics, education, etc. In fact, there is an important aspect of such analytics, that is, the detection and characterization of influential actors in social networks. Various studies have suggested different approaches and specific measures to solve the problem of influential actors detection.

Influential actors in social media have an effective role in information diffusion. For instance, A viral marketing operation for a new product can be conducted by

Ziyaad Qasem (e-mail: ziyaad.qasem@hs-ruhrwest.de)✉ · Marc Jansen
Computer Science Institute, University of Applied Science Ruhr West, Bottrop, Germany

Tobias Hecking · H.Ulrich Hoppe
Dept. of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, Duisburg, Germany

H. Cherifi et al. (eds.), *Complex Networks & Their Applications V*,
Studies in Computational Intelligence 693,
DOI 10.1007/978-3-319-50901-3_10

seeding the product in Twitter with a few elected influential actors who can influence others in a way that might help in the rapid spread of that product.

T measure [13] provides a new type of influence in online social network in order to emphasize on those actors who attract many outsiders to join the own community in which a specific topic is dealt. For example, in Twitter those actors spawn many retweets on a certain topic from people who have no previous contributions on that topic.

In this paper, the robust promise of influential actors detection leads us to extend T measure to present a new measure (HT measure) for the detection of influential actors which is based on quantifying the contribution of this actor to increasing the size of the network by attracting new attractors of the specific subcommunity. In other words, while T measure defines the attractiveness value of an actor through evaluating the number of outsiders who joined to the community by this actor, HT measure will refer to his/her attractiveness value through evaluating the importance of those outsiders. In the evaluation section of this paper, we apply our measure to a dataset based on Twitter communication around #EndTaizSiege (related to recent events in Yemen). We compare our measure with T, Katz centrality, indegree, and betweeness measures in terms of how good these measures are if used to refer to the influential actors in social media in terms to their ability to attract others to become active in the information diffusion process.

The rest of the paper is organized as follows: Section 2 presents related research. An overview of T measure approach is given in section 3, which also provides the basic formal definitions. Section 4 introduces the implementation of our measure, followed by the description of our datasets and the experimental results in section 5. Section 6 deals with the performance of our approach in the influence maximization problem. Finally, conclusions are drawn and an outlook for further research is described in section 7.

2 Related Works

Social influence analysis has attracted considerable research interests in recent years. A wide scheme of research focused on modelling and measuring influence and on influential actors detection. Particularly online social networks such as Twitter are of special interest. However, regarding the manifestation and identification there are still open questions.

It could be shown from the study presented by Cha et al. [2] that applying different measures can produce utterly different results when it comes to the task of ranking actors according to their influence. They illustrated an in-depth comparison of three measures of influence: indegree (number of followers of an actor), retweets (number of retweets containing ones actor name) and mentions (number of mentions containing ones actor name). They concluded that different measures can be used to identify different types of influential actors. Popular actors with high indegree were not necessarily influential in terms of spawning retweets or mentions and most influential actors can hold significant influence over a variety of topics. Consequently,

the way in which a network is extracted from social media content and the measure of influence should be considered carefully with respect to the roles and type of influence a one aims to reveal.

Qasem et al. [13] proposed a new approach which is related to the research presented in [2] in the sense that it aimed for a clear formulation of social influence and a methodology to produce an exact ranking of the actors according to the definition. In concrete, Qasem et al. [13] introduced a new type of influence in online social network to define those actors who attract many actors to join the own community in which a specific topic is dealt. Based on this type of influence, a new measure (T measure) has been proposed to define those actors.

In contrast to local measures that only take into account the direct neighbourhood of an actor, there exist also recursive measures that determine the centrality of an actor relative to the influence of its neighbours. A measure of influence proposed in the early years of social network analysis, which is still of importance, is the Katz centrality[7]. It accounts for the ability of an actor to spread information through a network by the counting the number of paths the actors have to each other actor. In addition, longer paths are weighted less than short paths.

Closely related measures are Eigenvector centrality for undirected networks and PageRank for directed networks. These measures are recursive in the sense that they calculate the centrality of each actor based on the centrality of its neighbours. Adaptations to Twitter a based on link analysis are TURank (Twitter User Rank) [16] utilizes ranking algorithm to present based on link analysis a new algorithm in which influential actors are defined. TURank defines actor-tweet graph where nodes are actors and tweets, and links are follow and retweet relationships. PageRank algorithm is extended by TwitterRank [15] to detect influential actors in Twitter based on link structure and topical similarity. Azaza et al. [1] proposed a new influence assessment approach depending on belief theory to combine different types of influence markers on Twitter such as retweets, mentions and replies. They used Twitter dataset of European Election 2014 and deduced the top influential candidates. These ideas were taken up in this work to assess the importance of an actor according to the potential to attract new actors to join the network. Here, the attraction value of an actor can be adjusted by the attraction values of the attracted actors achieve later on. In other words, high attractors are those who influence others to become active in the Twitter communication and also attract many others to do so.

Information diffusion in a network refers often to the influence in the spread of information. Particularly in social media, influential actors can control the diffusion of information through the network to some extent. Information diffusion is defined as the process by which a new knowledge or idea spread over the social networks by the means of communications among the social network actors [14]. The most widely used information diffusion models are the independent cascade (IC) [3][4] and the linear threshold (LT) [5]. These two models describe different aspects of influence diffusion. The IC and LT models have been introduced by Kempe et al. [8] to fix the problem of the influence maximization which search for those actors whose aggregated influence in the social network is maximized. whereas Pei et al. [12] provided strategies to search for spreaders based on the following of information

flow rather than simulating the spreading dynamics (modeled_dependent results). Furthermore, The features of identifying spreaders measures using independent interaction and threshold models through empirical diffusion data from LiveJournal are discussed in [11]. Morone et al. [10] proposed to map the problem of influence maximization in complex networks onto optimal percolation using CI (Collective Influence) algorithm.

Our work is related to the research presented in [13] in the sense that we aim to define a new type of influence based on the attractiveness model in order to detect those actors who attract new other attractors to participate the activities of the own community. As well as, our study is related to the approach of [7] in the sense that an actor is influential if he/she is linked from other influential actors. This new type of influence led us to propose a new measure (HT measure) to detect those actors, and compare the results with other standard measures. In this paper, we evaluated the performance of our measure in the information diffusion maximization problem by selected sets of top actors based on HT measure and other sets which are defined by T, Katz measure, and other standard measures.

3 Approach

The approach of T measure provides a new type of influence in online social network in order to emphasize on those actors who attract many outsiders to join the own community in which a specific topic is dealt. Thus, influential actors who are detected by T measure are those actors whose tweets spawn many retweets in a way that leads to an increase in the size of social network. T measure depends on the decomposition of a topical dataset that is collected from a social network according to the time period of collection. The basic idea of the dataset decomposition is to analyze a specific event in social media after each slice of time. The aim is to define the actors who affect the size of this event by attracting outsiders to participate. To be more specific, the attractiveness value (T value) of the actor A in the slice time t equals the number of new attractors who joined the community in the slice time $t+1$ by establishing new connection with actor A.

To formalize our HT measure, we will enumerate here briefly some of the concepts that are used to implement T measure.

The approach of T measure is based mainly on the decomposition of a topical dataset that is collected from a social network according to the time period of collection. This time period is referred to by the term P-period.

Definition 3.1 (P-period). P-period is a time duration of the data collection process from social networks.

The definition above is applied to the streaming dataset obtained from online social networks. If we have a historical dataset, P-period will be the period between the oldest activity (in Twitter, the activity would be tweet, retweet, reply, etc.) and the newest one in that dataset.

The social networks dataset in this approach is represented by a directed graph which is refereed to by P-graph.

Definition 3.2 (P-graph). P-graph is a directed graph constructed from social network data which have been collected during P-period.

Decomposition of a P-graph leads to decomposition of the P-period into slices of time so that every subgraph is related to a slice. This slice is referred by P-slice.

Definition 3.3 (P-slice). P-slice is a time slice of P-period.

If all P-slices are equidistant, the P-slice is called EP-slice.

Definition 3.4 (EP-slice). EP-slice is a P-slice in case all P-slices are equidistant.

To ease the definition of subgraphs of this approach, some terms related to actors according to P-slices are defined.

Definition 3.5 (P-actors). Let $s_1, s_2, \ldots s_n$ be the P-slices. For every i such that $0 < i \leq n$, the P-actors A_i is a set of all actors that joined the social network between the P-slices 0 and s_i.

Definition 3.6 (P_s-actors). Let $s_1, s_2, \ldots s_n$ be the P-slices. For every i such that $0 < i \leq n$, the P_s-actors A_{si} is a set of all actors that joined the social network between the P-slices s_{i-1} and s_i.

Figure 1 shows how the P-actors and P_s-actors are taken with respect to P-slice in this approach. The figure displays the P-actors A_3 and P_s-actors A_{s_3} as an example. A_3 is the set of all actors who joined the community until s_3 whereas A_{s_3} joined between P-slices s_2 and s_3.

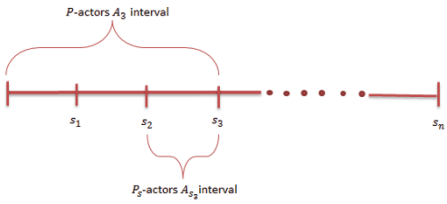

Fig. 1: P-actors and P_s-actors with respect to P-slices

The subgraphs used in this approach are defined as the following:

Definition 3.7 (P-subgraph). P-subgraph $G_i(A_i, E_i)$ is a directed subgraph of P-graph which is aggregated until P-slice s_i.

Definition 3.8 (S-subgraph). The i-th directed S-subgraph $S_i(A_i, E_{s_i})$ is the subgraph of the directed P-subgraph $G_i(A_i, E_i)$ with $E_{s_i} = \{(a,b) : (a,b \in A_{s_i}) \; or \; (b \in A_{i-1} \; and \; a \in A_{s_i})\} \cap E_i$

Figure 2 shows the difference between P-subgraph and S-subgraph in this approach where n is the number of P-slices and $1 < i \leq n$. P-subgraph G_{i-1} is the P-subgraph of the P-slice s_{i-1}, and P-subgraph G_i and S-subgraph S_i are of the P-slice s_i.

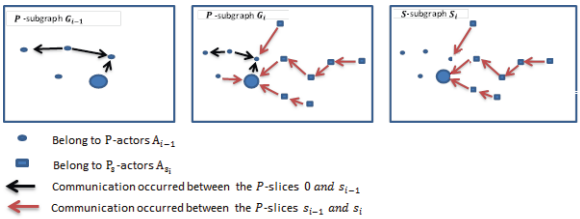

Fig. 2: Directed P-subgraphs G_{i-1} and G_i, and directed S-subgraph S_i

In the next section, we will introduce the implementation of our measure based on this approach.

4 Implementation

T measure tries to define those actors who attract many actors to the community. Figure 3 shows how the attractiveness value of the actor A is calculated with respect to T measure.

From figure Figure 3, T value of the actor A in the P-subgraph $G_{(i-1)}$ is equal to its indegree value in the $S-$subgraph S_i. Hence, The number of new actors joined the community by the actor A.

$$T(A_{G_{i-1}}) = indegree(A_{S_i}) \tag{1}$$

The indegree measure evaluates the number of neighbors of the actor A with order 1 (number of the immediate neighbors). In HT measure, we will increase the order to include the neighbors with order m, where m is the maximum neighborhood order. Thus, HT measure defines the attractors of attractors. Figure 4 shows the difference between T measure and HT measure.

From figure 4, HT value of the actor A in the P-subgraph $G_{(i-1)}$ is equal to its indegree plus the indegree of his/her neighbors with order m in the $S-$subgraph S_i.

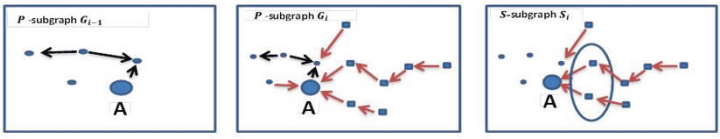

Fig. 3: T measure evaluation

$$HT(A_{G_{i-1}}) = T(A_{G_{i-1}}) + \sum_{a \in neighbors(A_{S_i}, m)} indegree(a_{S_i}) \quad (2)$$

Where m is the maximum neighborhood order.

HT and T values of the actor A in whole P-graph G are calculated as following:

$$T(A_G) = \sum_{i=1}^{n-1} T(A_{G_i}) \quad (3)$$

$$HT(A_G) = \sum_{1}^{n-1} HT(A_{G_i}) \quad (4)$$

Where n is the number of slices.

5 Evaluation

In this section, we will describe the evaluation strategy. Furthermore, the experimental results on the dataset will be discussed in this section.

5.1 Evaluation Strategy

We gathered a dataset from Twitter via Twitter API from December 31, 2015, to January 06, 2016. This Twitter dataset relates to the hashtag #EndTaizSiege (14,944 actors and 46,552 connections) that comprises a big connected component (containing 84% of actors), singletons (14%), and smaller components (2%).

Applying our approach leads to decompose P-graph constructed from Twitter dataset into three P-subgraphs and two S-subgraphs based on three P-slices. As a matter of fact, the time slicing has been estimated in accordance to the size of dataset using an equal window size for each slice. Figure 5 shows how the P-period with Twitter dataset #EndTaizSiege has been decomposed into equal window size so that we get a fair division of the retweet activities for each time slice.

The directed weighted P-graph of our collected Twitter dataset is constructed based on retweet activities so that actor a gets incoming connection from actor b if actor b retweeted a tweet of actor a. The weight of connection refers to the number of retweets between two connected actors.

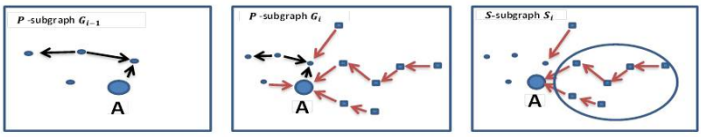

Fig. 4: HT measure evaluation

5.2 Experimental Results

For our Twitter dataset, we applied HT measure to verify whether it can detect influential actors. Table 1 shows the description of the top influential actors with respect to HT, T, Katz centrality, indegree, and betweenness measures. The question mark in the table 1 fields refers to an actor who is not a well-known as an influential actor within the community. We notice here how the HT and T measures refer to well-known influential actors within the community, or to the famous news accounts. Unlike other measures, the top ten influential actors with respect to HT and T measures are well-known within the community. In our case, the well-known actors have been recognized based on a local expertise, where they are the most renowned actors in the field of human rights and politics who continually traded their names in the newspapers and news concerning the current situation in Taiz city in Yemen. Their names have not been mentioned explicitly in order to protect their privacy.

Table 1: Description of top influential actors according to different influence measures in Twitter dataset #EndTaizSiege

Rank	HT	T	Indegree	Betweenness	Katz Centarlity
1	News Account N1	News Account N1	News Account N1	?	News Account N1
2	TV announcer T1	Journalist J1	Journalist J1	?	?
3	Journalist J1	TV announcer T1	TV announcer T1	?	Human Rights Activist H1
4	Human Rights Activist H1	Television reporter R1	Journalist R3	Journalist J2	Journalist J2
5	Human Rights Activist H2	Human Rights Activist H1	Human Rights Activist H1	?	?
6	Television reporter R1	Human Rights Activist H2	News Account N2	?	Television reporter R1
7	News Account N2	News Account N2	Human Rights Activist H2	Human Rights Activist H3	Journalist J1
8	Journalist J2	Political activist P1	?	TV announcer T1	TV announcer T1
9	Political activist P1	Journalist J2	Political activist P1	News Account N1	?
10	Political activist P2	Political activist P2	?	?	?

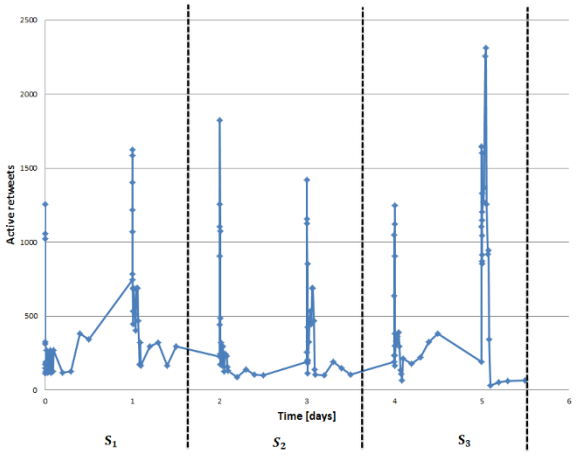

Fig. 5: Retweet activities over time in our Twitter dataset

6 Information Diffusion

In our work, we study the information diffusion to compare our measure with other existing measures in terms of how good these measures are if used to refer to the influential actors in social media in terms to their ability to attract others to become active in the information diffusion process. In order to assess how well the HT measure is suited to uncover influential actors with respect to information diffusion, we simulate the diffusion of information originating from a seed set of nodes through the Twitter networks using the well-known independent cascade (IC) model [8].

In information diffusion, the IC model is proposed where the information flows through cascade over the social network. In IC model, there are two terms are used to describe the state of the actors. The actor who is influenced by the information is called active, and inactive for the actor who is not influenced. The IC model process starts with activated actors as an initial seed set . In step s, an actor a will get a single chance to activate each currently inactive neighbor b. Actually, the activation process is based on the propagation probability P of the actors links. The propagation probability P of a link is the probability by which an actor can influence the other actors. In Twitter, we proposed that actor a is influenced by actor b if he/she retweeted from actor b in proportion to the tweets number of actor b. So, the propagation probability P in IC model is based in our Twitter dataset on the link weight divided by tweets number of target actor.

To compare the performance of actors sets selected by the HT measure with other influence measures, we selected sets of top actors based on the HT, T, and Katz centrality measures. As well as, we selected the sets identified by measures that are known to be good heuristics for seed set selection, namely degree and betweenness centrality [9].

6.1 Simulation of attraction processes with time-respecting paths

In this section,We will report results based on simulated attraction processes. To do so, we adapt the IC model that is known to simulate the diffusion of information through a network as described above. Information diffusion and attraction processes have some commonalities but differ on various aspects. In traditional information diffusion models such as the IC model, the network is usually considered as stable in the sense that the set of nodes and the set of edges do not change over time. However, the nodes changes their states ”inactive” and ”active” during the information diffusion process. Attraction, as it is studied in this paper is similar in the sense that actors who are not part of the community (i.e. do not have contributed a tweet) are inactive while others are considered as active. On the other hand, the original IC model does not account for the fact that the network grows when new actors become attracted to the community. Thus, the IC model was adapted to take into account the creation times of the edges. These time varying networks have special characteristics regarding reachability of node pairs since a walk on the graph can only take edges with increasing timestamp, which is known as the time-respecting property (see [6]). In this aspect, we added a new activation rule to the IC model which is: the actor who is activated in time t

cannot activate those actors who have been linked with him/her before the time t. To explain this activation rule in more details, we define the following terms:

Definition 6.1 (Path-time). The path-time of each link in the network is the P-slice number in which this link has been created.

Definition 6.2 (Activation-time). The activation-time of each activated actor is the path-time of the link by which this actor has been activated.

Now, we can state that the actor a can not activate the actor b if the link from b to a has a path-time later than the activation-time of the actor a.

Using this activation rule the simulation can be interpreted as an attraction process were actors who are already part of the communities can attract others only if their activity starts after the activator has become active.

Previous studies [13] have shown that a seed selection strategy based on indegree yields similar results as a selection strategy based on the T measure. This is also expected with respect to the high correlation between these two measures. However, the benefit of the T measure that distinguishes it from other measures is that time is explicitly taken into account. The experimental results in the next section support the assumption that the T and HT measure can identify important attractors in time varying networks while it boils down to indegree if time is neglected.

6.2 Experimental results

Here, we considered the dataset of #EndTaizSiege which is related to an organized event in Yemen. Hence, we got a highly connected component that is suitable for the application of our approach which is basically aimed to identify those actors who contribute to attract others to participate in a specific organized event. We simulated the information diffusion based on the IC model with time-respecting paths for seed sets of sizes $n = 1...25$ which are generated from different influence measures. The diagram in figure 6 shows the results of applying IC model on our Twitter dataset with different seed sets which identified by different influence measures. Comparing with other influence measure, we notice that the HT measure yield the best performance in information diffusion under the IC model with time-respecting paths for the seed sizes bigger than 11. Additionally, we statistically verified the results of simulation for each seed set using T-Test. In case of $n > 11$, the differences between HT and T measures are significant. For example, results for the seed set 12 show that there is a significant difference in the score of HT measure (M =1259.95; SD = 291.1128 conditions; t(19) = 3.678480757; P = 0.000). On the other hand, the differences among HT and indegree mearures are also significant in case of $n > 12$.

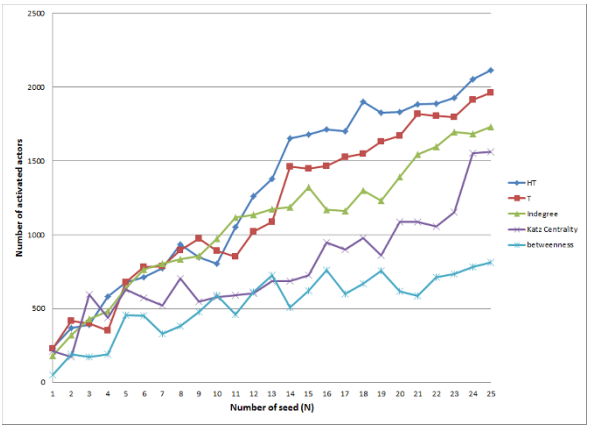

Fig. 6: IC model under time-respecting paths with different influence measures over Twitter dataset #EndTaizSiege

7 Conclusion

In summary, we presented in this paper an extended approach to detect influential actors based on the attractiveness model that is introduced with T measure. Our approach detects those actors who contribute effectively to increase the size of social network by attracting new attractors to the community in which a specific topic is dealt. Through experiment results we presented through how our proposed measure HT referred to the influential actors in Twitter dataset. Furthermore, we showed through experiment and statistical tests that the best performance has been yielded by HT measure in maximization of influence problem when we took the time into account.

Our current work in extending and improving this approach focuses on an elaboration of our measure with more datasets and more results, and describe it on multi-layer networks. Furthermore, we plan to develop an efficient general strategy for time slicing to determine the time period decomposition into time slices, and the role of time slicing in making HT measure far better than existing measures.

References

[1] Azaza, L., Kirkizov, S., Savonnet, M., Eric, L., Faiz, R.: Influence assessment in twitter multi-relational network. In: Proceedings of the 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 436–443 (2015)

[2] Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in twitter: The million follower fallacy. ICWSM **10**, 10–17 (2010)

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

[4] Goldenberg, J., Libai, B., Muller, E.: Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic

cellular automata. Academy of Marketing Science Review **9**, 1–18 (2001)
[5] Granovetter, M.: Threshold models of collective behavior. American journal of sociology pp. 1420–1443 (1978)
[6] Holme, P., Saramäki, J.: Temporal networks. Physics reports **519**, 97–125 (2012)
[7] Katz, L.: A new status index derived from sociometric analysis. Psychometrika **18**(1), 39–43 (1953)
[8] Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 137–146 (2003)
[9] Mochalova, A., Nanopoulos, A.: On the role of centrality in information diffusion in social networks. In: ECIS (2013)
[10] Morone, F., Makse, H.A.: Influence maximization in complex networks through optimal percolation. Nature (2015)
[11] Pei, S., Makse, H.A.: Spreading dynamics in complex networks. Journal of Statistical Mechanics: Theory and Experiment (2013)
[12] Pei, S., Muchnik, L., Andrade Jr, J.S., Zheng, Z., Makse, H.A.: Searching for superspreaders of information in real-world social media. Scientific reports (2014)
[13] Qasem, Z., Jansen, M., Hecking, T., Hoppe, H.U.: On the detection of influential actors in social media. In: Proceedings of the 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 421–427 (2015)
[14] Rogers, E.: Diffusion of Innovations, 5th Edition. Free Press, New York (2003)
[15] Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining, pp. 261–270 (2010)
[16] Yamaguchi, Y., Takahashi, T., Amagasa, T., Kitagawa, H.: Turank: Twitter user ranking based on user-tweet graph analysis. In: Web Information Systems Engineering–WISE 2010, pp. 240–253. Springer, Berlin Heidelberg (2010)