Optimal Detection of Influential Spreaders in Online Social Networks

Chee Wei Tan, Pei-Duo Yu, Chun-Kiu Lai, Wenyi Zhang and Hung-Lin Fu

***Abstract*—The wide availability of digital data in online social networks such as the Facebook offers an interesting question on finding the influential users based on the user interaction over time. An example is the clicking of the Facebook "*Like*" button to endorse a digital object (e.g., a post or picture) posted by other user. This online interaction activity connects users sharing similar opinions or disposition and spreads their influence. In this paper, we study the estimation problem of finding a small number of users in the online social network who are influential in maximizing the reach of a digital message when it originates from them. The digital interaction in the online social network can be modeled using an interaction graph, e.g., associate users through the past record of snapshot observations of *Like*'s activity in Facebook. We propose a network centrality approach in which we first use graph convexity to characterize the relative influential level of users on the interaction graph. We then propose a message passing algorithm to rank these users in order to identify the influential spreaders who play a forward-engineering role in catalyzing the spread of a new message. A useful application is to schedule a cascade of endorsement of a digital marketing message or for a business entity with a Facebook presence to find a number of Facebook users to spread the word of new commercial products. Lastly, we describe the performance of our algorithm using a synthetic dataset.**

I. Introduction

The wide availability of digital data in online social networks and the enormous user pool offers an interesting question on estimating the influence of users based on the user interaction over time. This online interaction over the online social network gives rise to a real-time interaction network that represents a fundamental medium for spreading and captures important characteristics on how information can diffuse. A prominent example is Facebook in which the digital contents (e.g., user status updates, posts, photos, videos, links) of a Facebook user are viewable on a *Facebook Timeline* by others who can interact with them (such as clicking the Facebook *Like* endorsement button for a post). These online interactions are recorded on the *Facebook Timeline* that again lead to more interaction. Here, the spreading process increases the susceptibility of other users to the same; this results in the successive spread of a digital message from a few users to many more. It is interesting to study the spreading impetus of a digital word-of-mouth engine starting from a selected few.

C. W. Tan (cheewtan@cityu.edu.hk), P.-D. Yu and C. K. Lai are with the City University of Hong Kong, W. Zhang is with The University of Science and Technology of China, and H.-L. Fu is with National Chiao Tung University. The research has been supported by the Doctoral Program of Higher Education (SRFDP) and Research Grants Council Earmarked Research Grants (RGC ERG) Joint Research Scheme through Specialized Research Fund 20133402140001, National Natural Science Foundation of China through Grant 61379003, the Research Grants Council of Hong Kong under Project M-CityU107/13 and Project 11212114.

It is reasonable to expect that a particular Facebook user who has a digital message in the past that has garnered many other Facebook users' interaction (say using the Facebook *Like* endorsement button) is *likely* to attract similar level of interaction with future posting of similar digital messages. This is because this digital interaction (e.g., the Facebook "*Like*") captures the desire to share similar opinions or disposition, and typically comes from Facebook users who are already socially close or displaying an interest to the particular Facebook user. Also, it captures the connectivity relationship among users in the online social network. This is useful such as when this particular Facebook user wants to schedule a cascade of endorsement for a digital marketing message or is a business entity who maintains a Facebook presence and wants to spread the word of new commercial products. By examining the past record on Facebook Timeline, this particular Facebook user can determine other Facebook users who are deemed influential enough in a viral marketing strategy [1].

In this paper, we study how to maximize the reach of a digital message in an online social network based on the past digital interactions. The motivation of this work is similar to prior work that analyze the effect of viral spreading and influence maximization [2]–[5]. The goal is to to find a small number of users (the seeding nodes in a graph) to spread a digital message. In this paper, Facebook users are modeled by nodes in a graph, and their associations of online interaction activity are modeled as edges in the graph. There can be many forms of digital interaction and we focus on the clicking of the Facebook *Like* button for endorsement. Snapshot records in the Facebook Timeline that capture this past digital interaction activity is used to find this group of users who then plays a forward-engineering role in catalyzing the spread to maximize the reach of a new digital message over time. When carefully selected, this initial number of users are the *influential spreaders* whose impact of influence are in turn recorded by the Facebook Timeline as past data that can be further reused (to refine the selection of future influential spreaders). This motivates a statistical inference of influential spreaders in catalyzing a viral spread.

The contributions of this paper are as follows:

- We propose a network centrality approach to the statistical inference of influential spreaders who are *likely* to maximize the spread of a digital object to as many users as possible in an online social network. This approach is motivated by the rumor source detection problem in which the center for a specially-constructed network centrality corresponds to a maximum-likelihood estimator for degree-regular tree graphs.

978-1-4673-9457-4/16/$31.00 ©2016 IEEE

- When the graph is a general tree, we provide a graph convexity characterization to this network centrality approach by showing that a good guess for this *most influential user* is equivalent to the graph-theoretic centroid of the tree. We give a result to ranking the users (nodes in the tree graph) in terms of the branch weight centrality.
- We propose a message passing algorithm to identify the influential spreaders by ranking the users when the tree graph is constructed based on some form of digital interaction using past data record. A heuristic that combines this with a breadth-first-search algorithm can address the general graph case, and we apply this heuristic algorithm to a synthetic dataset for preliminary evaluation.

This paper is organized as follows. In Section II, we give an overview of the work on a rumor source detection problem that specifically motivates the network centrality approach in this paper to identify influential spreaders. In Section III, we describe the basic problem and approach. In Section IV, we describe the graph convexity results and a message passing algorithm for ranking spreaders. In Section V, we evaluate the performance of our message passing algorithm on a synthetic dataset, and we conclude the paper in Section VI.

II. Related Work

Epidemic-like spreading is an important network science subject matter that has been extensively studied in the literature [2]–[7]. Indeed, the rampant spreading of malicious information has been identified as a major cyber-security challenge in networks [6]. For example, the spread of a computer virus in the Internet or a rumor in online social networks. This motivates the need to detect these information sources. Given a snapshot observation of the effect of spreading in the network (e.g., who possesses the malicious information and who is connected to whom), how to reliably identify the source of the spreading? This is a challenging problem that is complicated by the dynamics of the spreading and the problem size.

In the recent seminal work in [8], Shah and Zaman formulated this as a maximum likelihood estimation problem assuming a Susceptible-Infectious (SI) spreading model in [9]. A message is spread from a single initial node and thereafter the nodes that possess this message are called the *infected nodes* and those that do not are called the *susceptible nodes*. The authors in [8] introduced a *rumor centrality* to solve this problem exactly for degree-regular tree graphs. The rumor centrality is proportional to the likelihood of an infected node being the source and so the node with the maximum value (the *rumor center*) coincides with the optimal solution of a maximum likelihood estimation. In particular, this rumor center can be computed using message passing algorithms [10], and the detection performance can be quantified asymptotically, i.e., when the number of infected nodes becomes very large.

For general graph topology, solving this maximum likelihood estimation is still an open problem. Despite that, there are several suboptimal heuristics based on the rumor centrality that perform reasonably well (e.g., the breadth-first search heuristic in [8]). Since the work in [8], other related work include the generalization in [11] to random trees, the extension in [12] to incorporate suspect sets, extensions to multiple source detection and detection using multiple snapshot observations in [13] and [14], [15] respectively.

The authors in [16] provided a probabilistic characterization to the *rumor boundary* of the rumor spread data, and proposed a message passing algorithm to compute the likelihood for source estimation. The authors in [17] addressed the problem for the susceptible-infectious-susceptible spreading model demonstrating that heuristics based on the rumor centrality can work effectively. The authors in [18] provided an algebraic combinatorial analysis to characterize the asymptotic regime of the problem. The authors in [19] proposed a network centrality approach and a message passing algorithm to compute the harmonic influence centrality to measure the influence of nodes on the average opinion in networks.

III. A Network Centrality Approach to Inference

We assume the spreading to occur over an infinite network modeled as an undirected graph $G = (V, E)$, where $V = \{v_1, v_2, \dots\}$ is a countably infinite set of nodes and E is the set of edges of the form (i, j) for nodes v_i and v_j in V. In other words, the users in the online social network are the nodes in G, and the edges model the conduit for digital interaction. For example, two Facebook users are connected by an edge due to a Facebook Friend relationship or when they share similar opinions or disposition (such as one user endorsing the digital post of the other even when they are not Facebook Friend). As such, G is a digital interaction graph. The degree of a node v_i is the number of its neighbors denoted by d_i.

We assume a basic spreading model known as the Susceptible-Infectious (SI) model (e.g., see [9]) that is also used for the rumor detection problem in earlier work [8], [11], [12], [14], [15]. In this model, there are two types of nodes: (i) susceptible nodes that are capable of being infected (i.e., not yet possess the digital message); and (ii) infected nodes that can spread the digital message to their immediate neighbors. In this way, spreading occurs in a cascading manner, i.e., once a susceptible node receives the message from its neighbor, it retains the message forever and in turn may pass the message to its other susceptible neighbors, i.e., when $(i, j) \in E$. We also assume a memoryless property in spreading: let τ_{ij} be the spreading time for an infected node v_i to infect its susceptible neighbor v_j for all $(i, j) \in E$, then τ_{ij}'s are mutually independent and have exponential distribution with parameter λ (assume $\lambda = 1$).

A. Message Source Estimator

Let us suppose that the message originates from a node $v^\star \in V$ at a certain time $t = 0$ and spreads in the network G. Then, at time $t = T$, we observe the network G and find n infected nodes, which collectively constitutes a spread graph that we denote by G_n. Note that n represents the cardinality of the set of infected nodes in G_n. Obviously, the spread graph G_n is a connected subgraph of the underlying graph G as shown in Figure 1.

In the context of viral spreading, we have to choose the spreaders (say a single spreader $v^\star \in V$) at the outset $t = 0$

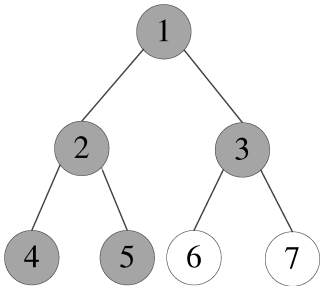

Fig. 1. An example tree network topology of a spread graph and its underlying graph. The grey and white circles in this figure represent respectively the infected nodes and susceptible nodes. Note that the underlying graph can be infinitely large, which is omitted due to page space.

without the hindsight of the spread graph G_n but with the intention that this G_n is reached in the shortest T for a given n (viral spreading means n to be typically large). As a first step, we may ask, had we known G_n at the outset from an oracle, then who is most likely to be the origin of the message, i.e., the spreader? This is simply akin to the rumor source detection in earlier work [8], [11], [12], [14], [15], and we may work backwards to find this spreader. The next step is then to leverage the structure of the estimator in the first step, without *a priori* knowing this spread graph G_n and relying only on past data to approximate one, to infer the spreaders.

In the following, we first focus on this first step (i.e., when the spread graph G_n is given) and examine the property of this estimator to identify a node $\hat{v}$ as the most likely spreader, assuming that G_n is a tree. From [8], the maximum-likelihood estimator that maximizes the correct detection probability is given by

$$\hat{v} \in \arg\max_{v \in Gn} P(G_n \,|\, v), \tag{1}$$

where $P(G_n \,|\, v)$ is the probability of observing G_n supposing that v is the original message source. Note that the solution of (1) may not be unique and so ties are broken uniformly at random. Solving (1) optimally is in general challenging. For G_n that are degree-regular tree graphs, the optimal solution to (1) is the rumor center [8], [11], [15].

IV. Graph Convexity Characterization

We now use graph convexity to provide an equivalent characterization to the rumor center in [8], [11], [12], [14], [15] for general tree graphs. For degree-regular tree special case, this means that, in the context of rumor source detection or the viral spreading in online social networks, the respectively *most likely rumor culprit* or the *most influential spreader* is equivalent to the centroid (see, e.g., [20]). This characterization is then extended to the relative ranking of nodes in a tree graph.

Let G_n be a rooted tree with the root at v_r where $v_r \in G_n$. For a tree graph G_n, finding the rumor center, i.e., $\max_i R(v_i, G_n)$ where $v_i \in G_n$, can be performed by a message passing algorithm that recursively computes the rumor centrality for each node starting from the leaves [8]. For any vertex v in the rooted tree G_n, a *parent* of v is its neighbor on the path connecting the vertex v and v_r. The children of v are its other neighbors, and we let $\mathsf{child}(v)$ denote the set of children nodes of the vertex v. If v is a leaf, $\mathsf{child}(v)$ is an empty set. A branch T_v^r of this rooted tree is a subtree with its root at v and we let t_v^r denote the order of T_v^r, i.e., the size of T_v^r in terms of the maximum number of children of v allowed.

Now, suppose that v_r is the message origin and the spreading has initiated, i.e., $G_1 = v_r$. Then, in G_2, this second infected vertex may be any child of v_r. Since there are $d(v)$ vertices in $\mathsf{child}(v)$ for any of this vertex, say u_i, where $i = 1, 2, 3, \ldots, d(v)$, we thus have [8]:

$$R(v, G_n) = \frac{(n-1)!}{t^v_{u_1}! \cdot t^v_{u_2}! \cdot \ldots \cdot t^v_{u_{d(v)}}!} \cdot \prod_{i=1}^{d(v)} R(u_i, T^v_{u_i}). \tag{2}$$

This can be expanded recursively from the root v_r to all the leaves of G_n to yield [8]:

$$R(v, G_n) = n! \cdot \prod_{u \in G_n} \frac{1}{t_u^v}. \tag{3}$$

Now, consider two adjacent vertices u and v in G_n and a vertex $w \in G_n - \{u, v\}$, then we have $t_u^v = n - t_v^u$ and $t_w^v = t_w^u$, where t_u^v is the order of a subtree T_u^v with v being the message origin and u as the root containing all the children of the tree. By using this recursion, it can be established that:

$$\frac{P(u|G_n)}{P(v|G_n)} = \frac{R(u, G_n)}{R(v, G_n)} = \frac{t_u^v}{n - t_v^u}, \tag{4}$$

which leads to the following result (see Proposition 1 in [8]).

Theorem 1: Given a tree G_n with n vertices, $v \in G_n$ is a rumor center if and only if

$$t_u^v \leq \frac{n}{2}$$

for all $u \in G_n - \{v\}$.

In words, this result characterizes the rumor center in terms of the sizes of its local subtrees.

We now introduce a graph-theoretic notion of G_n that provides an alternative characterization of the rumor center by using Theorem 1 as the link. Let us denote the branch weight of a local sub-tree of a vertex v in G_n by

$$\mathsf{weight}(v) = \max_{c \,\in\, \mathsf{child}(v)} t_c^v.$$

The vertex of G_n with the *minimum weight* is called the *centroid* of G_n [20]. By its definition, removing this centroid from G_n results in disconnected components in which the size of the biggest component is the smallest possible. Furthermore, the size of the smallest component is the biggest possible. For example, the centroid of the spread graph in Figure 1 is Node 2. Let us also define the *distance centrality* of $v \in G_n$ as $D(v, G_n) = \sum_{j \in G_n} d(v, j)$, where $d(v, j)$ is the distance (in terms of hop) between vertices v and j [8]. The vertex in G_n with the minimum distance centrality is called the distance center. We have the following result.

Theorem 2: Let G_n be a general tree graph and v is a vertex in G_n. Then, the following statements are equivalent:

1) The vertex v is a rumor center of G_n and also a distance center of G_n (proved in [8]).
2) The vertex v is a centroid of G_n.

It has been established in graph theory that a tree has either exactly one, or exactly two centroids joined by an edge (see, e.g., [20]). This implies that, by using Theorem 2, there are *at most two* rumor centers, and this scenario with two rumor centers happens only when the maximum branch size is exactly $n/2$. Furthermore, that the centroid and the distance center coincides has been pointed out in [20].

Now, a practical implication of Theorem 2 is that this rumor center in [8] for a given tree G_n can be found using alternative algorithms (such as those proposed in [20]–[22]) based on the notion of tree centroid (and to solve (1) optimally). All these alternative algorithms have computational time complexity $O(n)$ similarly to the message passing algorithm in [8].

A. A Message Passing Algorithm to Rank Centrality

Suppose the tree G_n is given, we give a message passing algorithm that ranks the users in terms of relative tree branch weight (or equivalently the rumor centrality). The ranking of the nodes makes use of the relative centrality measure between adjacent nodes and is equivalent in the sense of the rumor centrality, distance centrality or the branch weight centrality:

Theorem 3: Let G_n be a general tree with n vertices, and $u, v \in G_n$ are two adjacent nodes (neither u nor v needs to be the centroid). Then, the following statements are equivalent:

1) $R(v, G_n) \geq R(u, G_n)$.
2) $D(v, G_n) \leq D(u, G_n)$.
3) $\mathsf{weight}(v) \leq \mathsf{weight}(u)$.

Note that Theorem 3 implies Theorem 2. Another implication of Theorem 3 is that ranking can be determined using a message passing algorithm in which nodes exchange messages of computation in a recursive manner to determine this ranking. Suppose that all the nodes have a full knowledge of its distance to all the other nodes in the tree and its own degree. The network centrality computation can already be performed using a message passing algorithm such as that proposed in [8] or in [16]. As these messages are passed from the leaf nodes to their parent nodes who in turn aggregate the messages collected from their children nodes and pass the aggregated result to their parent nodes, this process iterates until all the nodes compute their network centrality.

In fact, a useful by-product of this recursive process is to find the top-k network centrality for a given input integer k as given in Algorithm 1 where we let K denote the set of nodes with top-k rumor centrality, and $\mathsf{CHS} = \{u | u \in \mathsf{child}(K)\}$. The output are the k influential spreaders. For any given integer k, Algorithm 1 has a computational complexity $O(k^2 \max_i d_i)$ to finding the top-k nodes, where $\max_i d_i$ is the maximum degree in G_n.

Algorithm 1 Algorithm for top-k Centrality Nodes

Input k, G_n with the centroid v_1 and branch weight centrality of each node
Set $K = \{v_1\}$, $\mathsf{CHS} = \phi$
for $i = 1, \ldots, k-1$ **do**
  $\mathsf{CHS} = \mathsf{CHS} + \{u | u \in \mathsf{child}(v_i)\}$
  set $v_{i+1} = \arg \min_{u \in \mathsf{CHS}} \{\mathsf{weight}(u)\}$
  $K = K + \{v_{i+1}\}$
end for

V. Breadth-First-Search Heuristic and Numerical Evaluation

Now, the spread graph, i.e., G_n, is unknown and so an interaction graph has to be constructed from past data. Also, in general, any interaction graph constructed using real data in online social networks is not a tree graph. We propose a top-k detection heuristic that first uses the breadth-first-search algorithm to obtain a breadth-first-search tree rooted at each user and then runs Algorithm 1 to find a number of highly-ranked (top-L with $L > k$) candidates for each rooted tree. The k candidates that appear the most frequently among all the rooted trees are then deemed to be the influential spreaders.

We use a synthetic dataset from *The Stanford GraphBase: A Platform for Combinatorial Computing* [23] that depicts the network of fictional human characters in Victor Hugo's 1862 novel Les Misérables. Each node of the interaction graph is a fictional character in Victor Hugo's 1862 novel Les Misérables, and there is an edge between two characters if they appear in the same chapter. There are some common features shared between the Les Misérables dataset and the digital interaction over a time-line. The Les Misérables dataset depicts human social interaction over time (by their appearance over the chapters). The novel consists of many chapters crisscrossed by a number of characters, and each chapter is relatively short (with few pages) and yet the correlation across chapters is overall rich enough to portray the key players (i.e., the protagonists).

We use this dataset as an initial evaluation of the effectiveness of our heuristic (in detecting the protagonists). The original interaction graph of seventy-seven characters for five volumes is shown in Figure 2. From this dataset, another subgraph is generated and this is the second interaction graph of thirty-seven characters of the first three volumes. We run the breadth-first search heuristic and Algorithm 1 using the two datasets and plot the partial results. Figures 3 and 4 show the top three candidates (shaded nodes) rooted at Valjean and Cosette respectively for the first dataset, and Figures 5 and 6 show the respective case for the second dataset. The top three protagonists identified from this methodology are the characters Valjean, Gavroche and Cosette.

VI. Conclusion

In this paper, we studied a basic problem of finding influential spreaders in online social networks for viral spreading. The goal was to maximize the reach of a digital message originating from some spreaders. This problem was broken down into two steps. The first step was to assume that the spread graph were supposedly known and the message origin to be estimated akin to the rumor source detection problem. The structure of this estimator was characterized using graph convexity to show that the centroid in graph theory is equivalent to the rumor center in the rumor source detection problem. This motivated a network centrality approach in the second step. We proposed to use past data records to

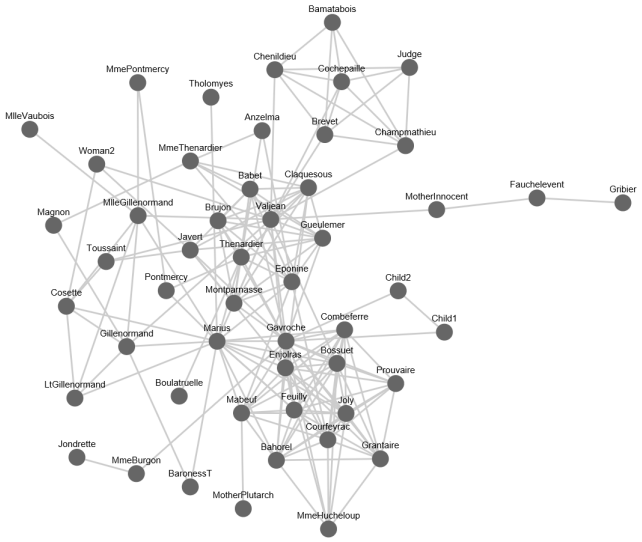

Fig. 2. The Les Misérables dataset from *The Stanford GraphBase: A Platform for Combinatorial Computing* [23] that depicts the network of fictional characters in Victor Hugo's 1862 novel Les Misérables.

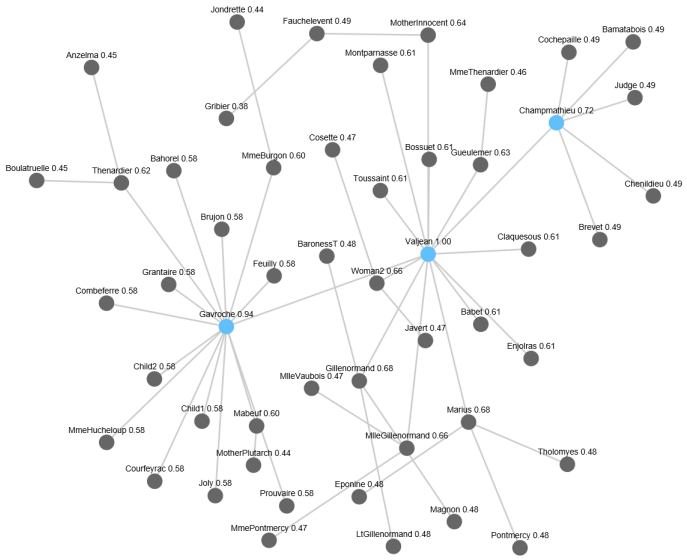

Fig. 3. Tree rooted at Valjean: the top-3 characters are Valjean, Gavroche and Champmathieu.

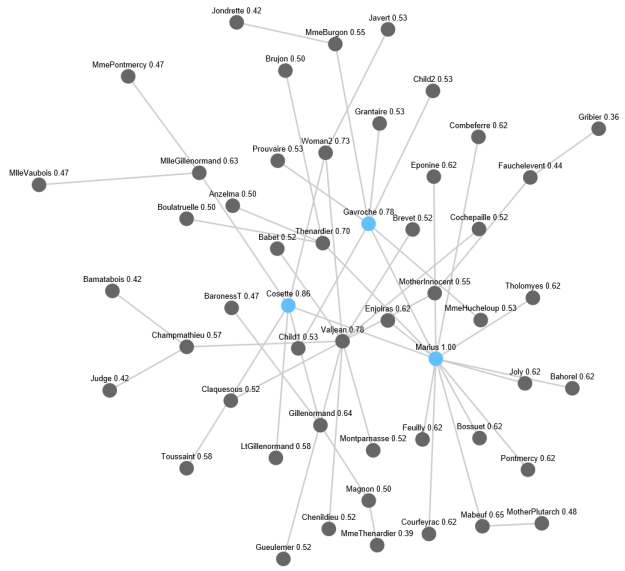

Fig. 4. Tree rooted at Cosette: the top-3 characters are Marius, Cosette, Gavroche.

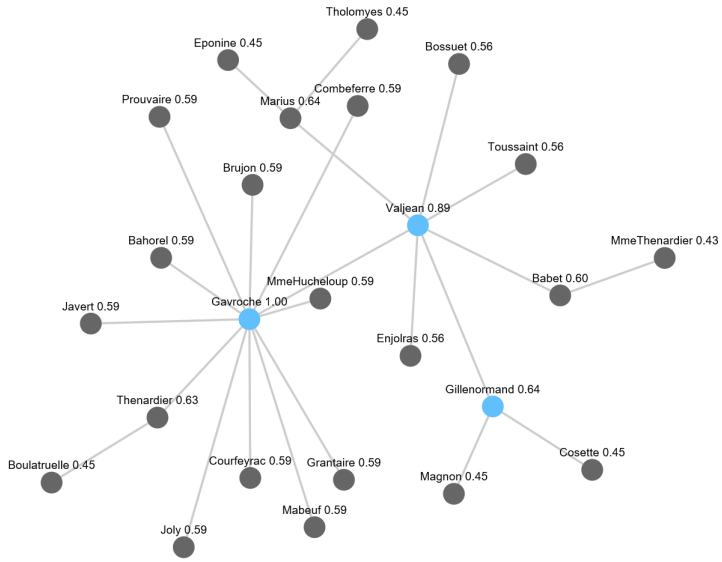

Fig. 5. Tree rooted at Valjean: the top-3 characters are Gavroche, Valjean, Gillenormand.

construct an interaction graph as an approximation to the spread graph. Examples can be snapshot observations recorded in the Facebook Timeline to capture the association of users' interaction (such as clicking the Facebook *Like* button). We proposed a message passing algorithm to rank the users so as to identify a number of spreaders that were deemed *most likely* to maximize the reach of a new digital message.

There are several future work. We are currently working on the software implementation using the Facebook Graph to evaluate the algorithm. The algorithm and software can be used to identify Facebook users who are most likely to *Like* a given digital object when it shows up on the Facebook Timeline. This can even lead to a useful approximation of the expected number of Facebook *Like*'s for this new digital message (useful for data analytics behind the newly-introduced Facebook's *Boost Like* marketing tool). In terms of modeling, we only looked at harnessing past records of Facebook *Like*'s activity in the Facebook Timeline to associate users in the online social network. How to harness past records and to refine the user association process in order to build an accurate interaction graph is important from a data analytics viewpoint. The network centrality of the interaction graph proposed in this paper is only a crude estimate of the influence in catalyzing a spread. It can be interesting to refine this approach by exploiting the correlation of user influence using different interaction graphs that are correlated over time.

APPENDIX

A. *Proof of Theorem 2*

Let G_n be a tree of size n and $v \in G_n$. Observe the following directions. Let us prove $(1 \Rightarrow 2)$: We prove it by

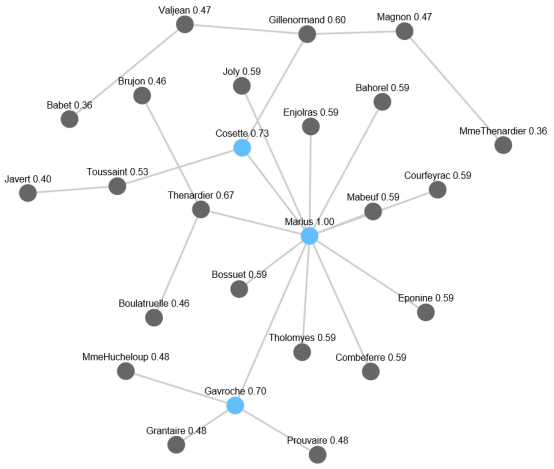

Fig. 6. Tree rooted at Cosette: the top-3 characters are Marius, Cosette, Gavroche.

contraposition argument. Suppose v is not a rumor center, by (2.2.1) there is a branch of v, say T_u^v, with order $> n/2$ and u is adjacent to v. Now, we need a relationship between $\sum_{s\in G_n} d(v,s)$ and $\sum_{s\in G_n} d(u,s)$ as described by

$$\sum_{s\in G_n} d(v,s) = \sum_{s\in G_n} d(u,s) + (t_u^v - 1) - (t_v^u - 1).$$

We have $\sum_{s\in G_n} d(v,s) > \sum_{s\in G_n} d(u,s)$, since $t_u^v > t_u^v$. This implies that v is not a distance center.

Next, let us prove (2 $\Rightarrow$ 3): First, we need the following fact: If all v's branches are of order $\leq n/2$, then v is the centroid. Again, by contraposition argument, suppose v is not a centroid, then there exists a branch of v whose order $> n/2$, that is, v is not a rumor center by (2.2.1).

Lastly, let us prove (3 $\Rightarrow$ 1): Suppose v is a centroid, then each of all its branches is of order $\leq n/2$. This implies that v is a rumor center. Let $u \in G_n$, if u is adjacent to v, then $\sum_{s\in G_n} d(v,s) < \sum_{s\in G_n} d(u,s)$ and we finish the proof. If u is not adjacent to v, then we can partition all the vertices in G_n into three sets. The first one is T_v^u, the second one is T_u^v and the last one contains all the vertices not in T_v^u and T_u^v, say R. Let l denote $d(u,v)$. Now, consider $\sum_{s\in G_n} d(v,s) - \sum_{s\in G_n} d(u,s) = (\sum_{s\in T_v^u} d(v,s) + \sum_{s\in T_u^v} d(v,s) + \sum_{s\in R} d(v,s)) - (\sum_{s\in T_v^u} d(u,s) + \sum_{s\in T_u^v} d(u,s) + \sum_{s\in R} d(u,s))$.

Since v is the rumor center, we have :
(1) $|R| + t_u^v \leq n/2$, and $t_v^u > n/2$;
(2) $(\sum_{s\in T_u^v} d(v,s) + \sum_{s\in T_v^u} d(v,s)) - (\sum_{s\in T_u^v} d(u,s) + \sum_{s\in T_v^u} d(u,s)) = l \cdot (t_u^v - t_v^u)$;
(3) $|\sum_{s\in R} d(v,s) - \sum_{s\in R} d(u,s)| \leq l \cdot |R|$.
Combining these three properties, we conclude that $\sum_{s\in G_n} d(v,s) - \sum_{s\in G_n} d(u,s) < 0$, for any $u \in G_n$, that is, v is the distance center. □

B. Proof of Theorem 3

Let G_n be a tree of size n, and $u, v \in G_n$. Observe the following directions. Let us prove (1 $\Rightarrow$ 2): Suppose $R(v,G_n) \geq R(u,G_n)$, we have $D(v,G_n) = D(u,G_n) - t_v^u + t_u^v$ and $t_v^u \geq t_u^v$, and so we conclude that $D(v,G_n) \leq D(u,G_n)$.

Next, let us prove (2 $\Rightarrow$ 3): Suppose $D(v,G_n) \leq D(u,G_n)$, we have $D(v,G_n) - D(u,G_n) = t_u^v - t_v^u \leq 0$. This implies that $t_u^v \leq t_v^u$. Note that $\mathsf{weight}(u) = t_v^u$. If not, then there is a branch of u with size larger than t_v^u thereby implying $t_u^v \geq t_v^u$, which is a contradiction. Hence, we have $\mathsf{weight}(u) = t_v^u \geq \mathsf{weight}(v)$.

Lastly, let us prove (3 $\Rightarrow$ 1): Suppose $\mathsf{weight}(v) \leq \mathsf{weight}(u)$, and note that $\mathsf{weight}(u) = t_v^u$. Since u is not the rumor center, we have $t_v^u > n/2$ and so $t_u^v \leq n/2$, this implies that $R(v,G_n) \geq R(u,G_n)$. □

REFERENCES

[1] A. Sela, I. Ben-Gal, A. S. Pentland, and E. Shmueli, "Improving information spread through a scheduled seeding approach," *Proc. of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2015.
[2] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," *Proc. of the ninth ACM SIGKDD international conference on knowledge discovery and data mining*, 2003.
[3] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physics Review E*, vol. 69, 2004.
[4] F. Chierichetti, J. Kleinberg, and A. Panconesi, "How to schedule a cascade in an arbitrary graph," *SIAM Journal on Computing*, vol. 43, no. 6, pp. 1906–1920, 2014.
[5] P. Shakarian, S. Eyre, and D. Paulo, "A scalable heuristic for viral marketing under the tipping model," *Social Network Analysis and Mining*, vol. 3, no. 4, pp. 1225–1248, 2013.
[6] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge Univ. Press, 1994.
[7] A. Ganesh, L. Massoulie, and D. Towsley, "The effect of network topology on the spread of epidemics," *Proc. IEEE INFOCOM*, 2005.
[8] D. Shah and T. Zaman, "Rumors in a network: who's the culprit?" *IEEE Trans. on Information Theory*, vol. 57, no. 8, pp. 5163–5181, 2011.
[9] N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases and its Applications*, 2nd ed. Griffin, 1975.
[10] D. J. C. Mackay, *Information Theory, Inference and Learning Algorithms*, 1st ed. Cambridge University Press, 2003.
[11] D. Shah and T. Zaman, "Rumor centrality: a universal source detector," *Proc. of 12th ACM SIGMETRICS/PERFORMANCE joint int. conf. on measurement and modeling of computer systems*, 2012.
[12] W. Dong, W. Zhang, and C. W. Tan, "Rooting out the rumor culprit from suspects," *Proc. of IEEE ISIT*, 2013.
[13] W. Luo, W. P. Tay, and M. Leng, "Identifying infection sources and regions in large networks," *IEEE Trans. on Signal Processing*, vol. 61, no. 11, pp. 2850–2865, 2013.
[14] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, "Rumor source detection with multiple observations: Fundamental limits and algorithms," *Proc. of ACM SIGMETRICS*, 2014.
[15] ——, "Rooting out rumor sources in online social networks: The value of diversity from multiple observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 663–677, 2015.
[16] L. Zheng and C. W. Tan, "A probabilistic characterization of the rumor graph boundary in rumor source detection," *Proc. of IEEE Digital Signal Processing*, 2015.
[17] Z. Wang, W. Zhang, and C. W. Tan, "On inferring rumor source for SIS model under multiple observations," *Proc. of IEEE Digital Signal Processing*, 2015.
[18] M. Fuchs and P.-D. Yu, "Rumor source detection for rumor spreading on random increasing trees," *Electronic Communications in Probability*, vol. 20, 2015.
[19] L. Vassio, F. Fagnani, P. Frasca, and A. Ozdaglar, "Message passing optimization of harmonic influence centrality," *IEEE Trans. on Control of Network Systems*, vol. 1, no. 1, p. 109120, 2014.
[20] B. Zelinka, "Medians and peripherians of trees," *Archivum Mathematicum*, vol. 4, no. 2, pp. 87–95, 1968.
[21] S. Zaks, "Optimal distributed algorithms for sorting and ranking," *IEEE Trans. on Communications*, vol. C-34, no. 4, p. 376379, 1985.
[22] O. Gerstel and S. Zaks, "A new characterization of tree medians with applications to distributed algorithms," *Networks*, vol. 24, no. 1, p. 2329, 1994.
[23] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing*. Addison-Wesley, Reading, MA, 1993.