Rumour Source Detection Using Game Theory

Minni Jain*, Aman Jaswani, Ankita Mehra, Laqshay Mudgal

Delhi Technological University, Delhi (India)

Received 25 February 2020 | Accepted 1 October 2020 | Published 22 October 2020

unir
LA UNIVERSIDAD EN INTERNET

Abstract

Social networks have become a critical part of our lives as they enable us to interact with a lot of people. These networks have become the main sources for creating, sharing and also extracting information regarding various subjects. But all this information may not be true and may contain a lot of unverified rumours that have the potential of spreading incorrect information to the masses, which may even lead to situations of widespread panic. Thus, it is of great importance to identify those nodes and edges that play a crucial role in a network in order to find the most influential sources of rumour spreading. Generally, the basic idea is to classify the nodes and edges in a network with the highest criticality. Most of the existing work regarding the same focuses on using simple centrality measures which focus on the individual contribution of a node in a network. Game-theoretic approaches such as Shapley Value (SV) algorithms suggest that individual marginal contribution should be measured for a given player as the weighted average marginal increase in the yield of any coalition that this player might join. For our experiment, we have played five SV-based games to find the top 10 most influential nodes on three network datasets (Enron, USAir97 and Les Misérables). We have compared our results to the ones obtained by using primitive centrality measures. Our results show that SV-based approach is better at understanding the marginal contribution, and therefore the actual influence, of each node to the entire network.

Keywords

Game-Theory, Jaccard Similarity Coefficient, Network Centrality, Rumour Source Detection (RSD), Shapley Value (SV).

DOI: 10.9781/ijimai.2020.10.003

I. Introduction

RUMOUR Source Detection (RSD) aims to identify the most powerful nodes that are the primary sources of rumour propagation within a network. Social networking has become a modern tool for people to connect and spread the news with the development of science and technology. Diffusion of information in a social network can occur at lightning speeds and more often than not, this is considered a boon when it comes to relevant and correct information being spread. But at the same time, these networks can also be used to spread false or unverified information, either deliberately or by mistake. Therefore, rumours spread quickly and widely, and they have a great power of destruction. It is therefore of great theoretical and practical importance to decide whether there is an influential spreader and to recognize who is the influential spreader in the process for prevention and control of rumour propagation. This task is considered to be challenging due to the high speed of diffusion of information, and also because of the continuously evolving and dynamic nature of these social networks.

The most common approaches to finding the most influential node used in the past include single centrality [1] and group centrality [2] measures. The four major centrality measures are as follows. First, Degree Centrality (DC) refers to the number of associations that a node has with other nodes in a network. For an undirected graph, it is taken equal to the number of nodes to which a node is directly connected. For a directed graph, we need to compute the in-degree as well as the out-degree for each node. Second, Eigen-Vector Centrality (EVC) considers the relative power or significance of the nodes. Here, each node is assigned a value representing its relative significance considering the fact that nodes which are connected to high-power nodes have a stronger influence over the network in comparison to those which are connected to low-power nodes. Third, Betweenness Centrality (BC) measures how strongly two nodes are connected via a given node. It is estimated as the ratio of the aggregate of shortest distances between any two nodes in the network, on which the node lies, to the shortest path between the two nodes considered. Finally, Closeness Centrality (CC) measures how quickly rumour can be spread from one node to all the other nodes in a network. It is measured as the inverse of the total sum of all shortest path distances between a given node and all other nodes in a network. For more insight into centrality measures along with mathematical derivations, refer to [3].

But these measures have a lot of disadvantages as different measures are based on different concepts and emphasize upon different topological properties of the network. For instance, DC gives the same weight to all the neighbours of a node when computing its importance. It would be more intuitive to give higher weights to nodes that are themselves important. In EVC, most of the weights get concentrated in a relatively smaller subgraph and therefore, all nodes are not quantified as they should be [4]. The remaining measures do not tend to capture the flow of information in the graph. Moreover, single centrality measures suffer from an inevitable disadvantage due to the failure to recognize the effects when considered in groups on node functionality. Group centrality measures were created to overcome this barrier and place great focus on operating in groups of nodes and not on their individual functionalities. Nonetheless, group centrality also suffers from a drawback as it focuses on a-priori-determined node groups and contributes to confusion when prioritizing individual nodes within the network.

* Corresponding author.
E-mail address: minnijain@dtu.ac.in

We aimed to work on game-theoretical algorithms to explore different strategies and metrics to assess the root cause of the rumour spread. Game-Theory is a significant paradigm that finds its applications in various fields. It is used in statistics and business analytics for prototyping the interactivity among participating agents [5]. Game-Theory has helped us to improve our presentiment, allowing for a logical analysis of various ideas which can be implemented in tandem with decision theory. Game-Theory has been widely used in the field of natural language processing. One of its most prominent applications is finding the most influential node within a network, which is relevant to our problem statement. We also do not face any of the above-mentioned disadvantages in this approach. Typical social network analysis cannot capture the dynamics of strategic interactions among the individuals in the network. Our proposed model is based on cooperative game-theory that solves this issue [6]. The elemental constituents of intricate interactivities in a network can be efficiently processed using a rich class of games, called influence games, as has been demonstrated in [7].

Shapley Value (SV) algorithm is a game-theoretic approach that has been explored in the past for finding the most influential nodes in a graphical network [8], [9], [10], but not for RSD problem specifically. The strategic issues in the Gale-Shapley model and its applications have been discussed in [11]. On the basis of the concept of marginal (or borderline) contribution, an important solution concept was proposed. Player i's SV, denoted by $SV_i(v)$, is equal to the weighted mean of i's borderline contributions to each coalition C, to which the player may belong.

$$SV_i(v) = \frac{1}{n!}\sum_{\pi\in\pi(I)}\{v(C_\pi(i)\cup\{i\}) - v(C_\pi(i))\} \quad (1)$$

In (1), the aggregate count of players is given by 'n' while $\pi(I)$ gives the set of all permutations with 'n' players. This concept is based on cooperative game-theory - an aspect of game-theory which encourages players to form coalitions to maximize their yield in the game. Coalitions are gatherings of players that form the essential or fundamental elements of decision making. These are assumed to uphold cooperative conduct which makes it reasonable to view these games as a contest between alliances of participants and not between separate players. The core assumption here is that as the game proceeds, an eminent alliance or coalition comprising all participants will manifest eventually. The theory of cooperative games provides a high-level approach as it describes only the coalitions' structure, strategies and benefits. More insight into the SV algorithm and its derivation can be found in [8].

We have used SV-based centrality algorithm that is based on the key idea of a game-theoretic network which means defining a cooperative game across a network in which agents are nodes, coalitions are node groups, and coalition payoffs are defined to meet the requirements of a given application. The main contribution of our work is that we have explored the power of five different variations of the SV algorithm on various social networks that can be used for the purpose of spreading rumours.

We also used main centrality measures to identify the prominent top-k nodes to demonstrate a distinct and detailed contrast between our game-theoretical approach and the measures of prime centrality. Such a good analogy helped to portray the game-theoretical algorithm's aspects and accuracy vividly.

Section II gives a detailed study of various works done in the related field. Section III explains the datasets used and the algorithmic flow. Section IV describes the results obtained and the evaluations performed. Section V discusses the results and gives a theoretical explanation for the obtained results. Section VI concludes the research work with an insight into its future scope.

II. Related Work

One of the fundamental research discussions in the literature on network analysis is the topic of connectivity. The first to experiment to detect the primary top-k nodes were Domingos and Richardson [12]. They developed an algorithmic model to address this problem by modelling social media network as Markov random fields which mathematically characterized the probability of occurrence of an event.

Chen and Teng [1] explained that single node centrality measures are suitable for assessing individual influence in isolation while Shapley centrality assesses individuals' performance in group influence settings. Wei et. al. [2] explored the need to learn distributed vector representation for each vertex in a network. They laid emphasis on node classification and link prediction. An interesting approach to discover influential nodes in a network by formulating a target set selection problem has been discussed in [9]. Here, the problem comprises two main steps – the first step deals with finding a set of 'k' key nodes that would maximize the number of nodes being influenced in the network, while the second step is based on the λ-coverage problem.

We further investigated various kinds of centrality measures used for finding the most influential nodes in a network. DC, discussed by Gao et. al. [13], is used to efficiently measure the significance of nodes. However, it suffers from a severe disadvantage which is that it does not take into consideration the overall, detailed anatomy of the network. EVC, according to Stephenson and Zelen [14], overcomes the defects associated with DC. It takes into account the influence of neighbours of the node in consideration. BC, as explored by Freeman [15], learns topology-related data of networks in advance. Al-Garadi et. al. [16] describes how CC can be efficiently used to identify multiple influential spreaders. We also investigated the disadvantages associated with using centrality measures to find the most influential node in [1], [17], [18], which have been discussed in section I.

An attempt has been made to find the most influential node in a network using mapping entropy (ME) that reflects the correlation between a node and its neighbours [18]. We particularly inspected the application of ME using ENRON email dataset which is commonly used for the study of social networks [19]. ME recognizes the significance of a node in a complex network based on the knowledge of degree of the node and degrees of its neighbours. This technique for network attack helps to identify the node to attack, thereby saving valuable resources. However, the game-theoretic approach, that has been proposed, is able to capture and take into account the interactivity and dynamics of strategic interactions in a network, not only with immediate neighbours, but also with a larger subset of relationships in the graphs. Thus, we chose an SV-based algorithm to find the most influential nodes in a cooperative game.

Previous research by Tan et. al. [20] on spreading rumours focused primarily on communities' viral epidemics. The normal (and somewhat standard) model for viral epidemics is called the restored or SIR model that is susceptible-infected- recovered. There are three types of nodes in a typical rumour propagation model: i) vulnerable nodes capable of infection, ii) infected nodes capable of further spreading the virus, and iii) recovered nodes that are healed and no longer capable of infection. The most influential spreaders of rumour are identified. Various methods have been defined for the same including weighted k-core decomposition method [15] and rumour centrality with a mass centre technique [20]. An advanced form of this model, called the SEIR model, was also studied. Zhou et. al. [21] considered the graph topology and observed snapshots in a network to identify the single rumour source by formulating the nodes in a network into four possible states: susceptible (S), exposed (E), infected (I), and recovered (R).

We studied about Explosion-Trust (ET) Game Model by referring to [22]. It remarkably explains how a rumour spreading model can be constructed using game-theory by considering two very significant factors – rumour explosion degree and trust degree of the source node. In [23], a unique Belief-Propagation-based (BP) algorithm has been discussed that computes the joint likelihood function of the source location and the spreading time for the general continuous-time to detect the rumour source in a network. In [24], the concept of influence maximization has been explained from a game-theoretical perspective. A Coordination Game (CG) model, in which every individual node makes its decision based on the benefit of coordination with its network neighbours, has been proposed. SV or other game-theory solution theories can be applied to other network-related issues as well, for example, to the cost allocation problem in the electric market transmission system, and for each application, the mathematical aspects of the problem should then be addressed.

The original SV algorithms that have been implemented using Monte-Carlo simulations in the past are computationally expensive and may not arrive at an exact answer. Michalak et. al. [8] developed approximate analytical formulas for these simulations that run in polynomial time. They discuss five characteristic functions, each of which tries to convey a certain centrality concept. We have taken inspiration from their work and worked with five SV games that focus upon one characteristic function each. Furthermore, to show the comparisons of our work with existing literature, we have taken the works of Qiao et. al. [25], Hardin et. al. [26] and Munjal et. al. [27]. We found very few works that list out the top 10 most influential nodes on one of the datasets that we used in our study, with the help of primitive centrality measures. Hence, we have used these three works for our comparative study. Qiao et. al. [26] explored an entropy-based centrality measure along with the primitive centrality measures and tested it on the USAir97 dataset [28]. Hardin et. al. [26] studied the relationships in the Enron dataset [29], [30] using six centrality measures. Finally, Munjal et. al. [27] found the most influential nodes from the Les Misérables dataset [31]. We have performed our experiments on these three datasets and compared the top 10 most influential nodes obtained by using our five SV games, with the top nodes listed in these works. More details about the datasets used are given in section III.A.

III. Proposed Method For RSD

Section III.A explains the datasets used and their importance. Section III.B explains the algorithmic flow used in detail.

A. Datasets

This section gives an elaborate description of the datasets that have been used for our implementation. For our experiments, we required undirected, positive weighted-graphs that could be expressed as social networks, the top 10 most influential nodes of which were already known (so that we could compare our results with these already known influential nodes). We have used three major datasets which satisfy these criteria and they have been described below.

Unweighted Graph

An unweighted graph can be technically defined as a graph G(N, E) having 'n' nodes represented by set N and 'e' edges represented by set E consisting of unordered pairs, such that $(n_1, n_2) = (n_2, n_1)$ and $(n_1, n_2) \in E$ and $n_1, n_2 \in N$. Games 1 and 2 are played by creating an unweighted network from the datasets.

Weighted Graph

A weighted graph can be technically defined as a graph G(N, E) having 'n' nodes represented by set N and 'e' edges represented by set E consisting of ordered pairs, such that $(n_1, n_2) \neq (n_2, n_1)$ and $(n_1, n_2) \in E$ and $n_1, n_2 \in N$. Games 3 – 5 are played by creating a weighted network from the datasets.

1. Enron Dataset

The CALO Project (A Cognitive Assistant that Learns and Organizes) compiled and planned this dataset [29], [30]. It contains data from about 150 users, belonging to the Enron organization, grouped into files, mainly senior Enron executives. There are a total of about 0.5 M messages in the corpus. We used a subset of this dataset, containing 143 nodes (people from the Enron organization) and 1800 edges (an edge exists between two people if they have communicated with each other via email). Edges are weighted with the frequency of email exchanges between two users. This dataset can act as a social network which can be used to spread rumours within the members of the organization. Hence, we can identify the important nodes and assign them labels that symbolize their relative network value. This dataset has been commonly used for the study of social networks as well as for finding the most influential nodes [19], [26], [32], and so we have compared the results of our algorithm with other studies that used the same dataset [26].

2. Les Misérables

This is a co-occurrence graph for the characters that appear in the novel 'Les Misérables' by Victor Hugo [31]. This dataset consists of 77 nodes and 254 edges where a node represents a character and an edge between two nodes shows that these two characters appear in the same chapter of the book. The weight of each link indicates how often such a co-appearance occurs. This dataset too can act as a social network for the spread of a rumour. We have compared the results obtained by our SV-based approach with those obtained by various centrality measures used in [27].

3. USAir97

USAir97 dataset [28] has been transformed into an undirected network, created by 332 nodes, where one airport represents a node, and 2126 edges, with each edge reflecting a direct airline between two American airports if any. Here, weights represent the normalized distance between two airports. This dataset is not particularly useful for the purpose of rumour spreading but due to lack of supervised datasets with their most influential nodes known to us, we have included this dataset to test the results of our approach with the most influential nodes obtained by various centrality measures, as in [25].

B. Algorithm

Focusing on Game-Theory's Shapley algorithm, we referred to the algorithms described in Michalak and Szczepański's work [8]. In both weighted and unweighted networks, the exact analytical formulae for SV-based centrality were established. The SV-based centrality polynomial-time algorithms have been developed.

1. Creation of Weighted and Un-Weighted Network Graphs

Graphs were created by using the *networkx* library in Python for all three datasets. Games 1 and 2 require unweighted graphs whereas the remaining games require weighted graphs.

2. Coalition Games Based on Shapley Algorithm

SV is the average marginal cost contribution across all potential coalitions of the function value. The Shapley algorithm was applied carefully and it tries to find the top-k nodes that might be the most prominent nodes.

Specifically, we concentrated on five underlying network-defined coalition games that vary in degree and centrality of the network. Each game has a certain characteristic function v(C) which represents how prominent a particular node is to a given coalition C.

For more insight into the working of these games and their underlying mathematics, refer to [8]. The game descriptions are as follows:

a) Game 1: In this game, we considered all the permutations of all the nodes that are immediately reachable, by one hop to the node $n_i \in$ N(G). Let each random permutation be denoted by P_i, the neighbours of node n_i, in the graph G(N, E) be denoted by n_i.neighbours and the degree of node n_i, be denoted by n_i.degree.

Algorithm 1 describes the procedure involved in SV calculation.

Algorithm 1: SVs for Game 1
Input: An unweighted graph G(N, E)
Output: SVs of all nodes in G
Initialise: $\forall\, n_i \in$ N(G) set SV[n_i] ← 0
for each $n_i \in$ N(G) **do**
  SV[n_i] ⟵ 1/(1 + n_i.degree)
  for each $u_i \in n_i$.neighbours **do**
  $SV[n_i] \leftarrow SV[n_i] + \frac{1}{(1+u_i.\text{degree})}$
  end for
end for
return SV

b) Game 2: In many real-life social scenarios, often taking into account nodes that are directly attached to each other is not enough. A rumour source will, more often than not, affect farther nodes.

For the purpose of taking relationships with farther nodes into account, and generalising the game, we introduced a value, p, depicting the number of agents that the node is adjacent to in a coalition. In this game, a node is considered 'influenced' if at least p of its neighbours are influenced. We divided the analysis using this game into two parts, first, where the degree of the node is less than p and second where the degree is more than p.

Algorithm 2 describes the procedure involved in SV calculation.

Algorithm 2: SVs for Game 2
Input: An unweighted graph G(N, E) and a positive integer p
Output: SVs of all nodes in G
Initialise: $\forall\, ni \in$ N(G) set SV[ni] ← 0
for each $ni \in$ N(G) **do**
  $SV[ni] \leftarrow \min\left(1, \frac{p}{(1+n_i.\text{degree})}\right)$
  for each $ui \in ni$.neighbours **do**
  $SV[n_i] \leftarrow SV[n_i] + \max\left(0, \frac{u_i.\text{degree} - p + 1}{u_i.\text{degree} * (1+u_i.\text{degree})}\right)$
  end for
end for
return SV

c) Game 3: In this game, we introduced the concept of weighted graph networks. This game is an extension of game 1; it uses the Dijkstra Algorithm to compute the distance between 2 nodes. The cutoff value, d, is the maximum permissible distance of a node from any member in a given coalition.

The extended degree is defined as the size of the set of all nodes that are at most distance 'd' away from the node n_i.

Algorithm 3 describes the procedure involved in SV calculation.

Algorithm 3: SVs for Game 3
Input: A weighted graph G(N, E, W) and a positive cut-off value d
Output: SVs of all nodes in G
Initialise: $\forall\, n_i \in$ N(G) set SV[n_i] ← 0
for each $n_i \in$ N(G) **do**
  Distance_Vector D ← Dijkstra(n_i)
  extended_neighbours ← empty 2D array
  extended_degree[n_i] ← 0
  for each $u_i \in$ N(G) such that $u_i \neq n_i$ **do**
    if D[u_i] ≤ d **then**
    extended_neighbours[n_i].add(u_i)
    extended_degree[n_i]++
    end if
  end for
end for
for each $n_i \in$ N(G) **do**
  $SV[n_i] \leftarrow \frac{1}{1+\text{extended\_degree}[n_i]}$
  for each $u_i \in$ extended_neighbours[n_i] **do**
  $SV[n_i] \leftarrow SV[n_i] + \frac{1}{(1+\text{extended\_degree}[u_i])}$
  end for
end for
return SV

d) Game 4: This is a generalization of game 3. Here we worked with the assumption that a node closer to a coalition will have a greater effect on it than some other node farther away, even if both

nodes satisfy the cut-off criteria as in game 3.

For this purpose, we introduced a positive-valued decreasing function f(x). f(d) refers to the function which has a directly proportional effect on SV of the coalition which is 'd' units away from a node.

The marginal contribution of each node n_i through node $n_i \neq n_j$, for each coalition C_i gives SV, as shown in Algorithm 4.

Algorithm 4: SVs for Game 4
Input: A weighted graph G(N,E,W) and function $f: \mathbb{R}^+ \rightarrow \mathbb{R}^+$
Output: SVs of all nodes in G
Initialise: $\forall\, n_i \in$ N(G) set SV[n_i] ← 0
for each $n_i \in$ N(G) **do**
  [Distance D , Nodes w] ← Dijkstra(n_i)
  sum ← 0, index ← |N|-1 , prev_dist ← -1, prevSV ← -1
  while index > 0 **do**
    if D(index) == prev_dist **then**
      currSV = prevSV
    else
      $\text{currSV} = \frac{f(D(index))}{1+index} - \text{sum}$
    end if
    SV[w(index)] ⟵ currSV + SV[w(index)]
    $\text{sum} \leftarrow \text{sum} + \frac{f(D(\text{index}))}{1+\text{index}}$
    prev_dist = D(index) , prevSV = currSV
    index ⟵ index - 1
  end while
  SV[n_i] ⟵ SV[n_i] + *f(0)* – sum
end for
return SV

e) Game 5: This is a generalization of game 2 in case of weighted networks. Here, we have defined a cut-off value (n_i) for each $n_i \in N(G)$. $d(n_j, C) = \Sigma_{nj \in ni.\text{neighbours}} W(n_i, n_j)$ for every coalition C, where $W(n_i, n_j)$ is the weight of the edge between nodes n_i *and* n_j (0 if no edge exists).

A node n_i marginally contributes node $n_j \in n_i$. neighbours to the value of coalition C_i if and only if $n_j \notin C_i$ and $d(n_j) - W(n_i, n_j) \leq W(C_i, n_j) < d(n_j, C)$.

Algorithm 5 describes the procedure for calculating the SVs.

Algorithm 5: SVs for Game 5

Input: A weighted graph G(N, E, W) and cut-offs $Wcutoff(n_i)$ for each $n_i \in N(G)$

Output: SVs of all nodes in G

Initialise: $\forall\, n_i \in N(G)$ set $SV[n_i] \leftarrow 0$

for each $n_i \in N(G)$ **do**

compute and store α_i and β_i

end for

for each $n_i \in N(G)$ **do**

for each m in 0 to $n_i.degree$ **do**

compute $\mu \leftarrow \mu(X^{ii}_{m})$, $\sigma \leftarrow \sigma(X^{ii}_{m})$

compute $p \leftarrow \Pr\{\mathcal{N}(\mu, \sigma^2) < Wcutoff(n_i)\}$

$SV[n_i] \leftarrow SV[n_i] + \frac{p}{1 + n_i.degree}$

end for

for each $v_j \in n_i.neighbours\ n_i.neighbours$ **do**

p ← 0

for each m in 0 to $n_i.degree$ **do**

compute $\mu \leftarrow \mu(X^{ij}_{m})$, $\sigma \leftarrow \sigma(X^{ij}_{m})$

compute $z \leftarrow \mathcal{Z}^{i}_{m}$

$p \leftarrow p + \frac{z * (v_i.\text{degree-m})}{v_j.degree * (v_j.degree+1)}$

end for

$SV[n_i] \leftarrow SV[n_i] + p$

end for

end for

return SV

3. Estimating Centrality Measures

After working on the five coalition games, we introduced multiple centrality measures to determine the network's most powerful node with the highest scope or effect. To generate an elaborate comparison, various network centrality measures such as DC, EVC, BC, CC, have been used.

IV. Results

We experimented on three real-world network datasets - USAir97 dataset [28], Enron email dataset [29], [30] and Les Misérables dataset [31], and then compared the results of five coalition games defined previously, with the results obtained using the four aforementioned centrality measures. Qiao et. al. [25] has applied these centrality algorithms using the USAir97 network to assess the performance of network centrality model. Table I accurately shows for USAir97 dataset, the comparison between the top-k (k=10) nodes identified by our model for all the five coalition games and those identified by various centrality models employed in [25]. Also, Table II shows for Les Misérables dataset, the comparison between the top-k (k=10) nodes identified by our model for all the five coalition games, and those identified by various centrality models employed in [25].

We observed that the number of common items between the top-10 nodes found using coalitional game 1 and those found using DC, BC, and CC measures are nine, nine and four, respectively. The most significant observation is that the top-10 nodes are the same for both the coalitional game 1 and EVC measure. For Les Misérables dataset, we observed that node 11 was recognized as the most influential node in all five coalitional games and also using DC, BC, and CC measures. We noticed an overlap of six nodes in the observations of game 3, game 5 and CC measure.

TABLE I. Comparison Between Shapley and Centrality Values Using USAIR97 Dataset

DC	BC	CC	EVC	Proposed Model				
				Game 1	Game 2	Game 3	Game 4	Game 5
118	118	118	118	118	261	261	118	118
261	8	261	261	261	118	118	261	261
255	261	67	255	255	152	152	182	67
182	201	255	182	166	182	182	152	255
152	47	201	152	152	255	255	201	201
230	182	182	230	182	230	230	255	166
166	255	47	112	230	201	201	230	293
67	152	248	67	67	8	8	8	248
112	313	166	166	147	166	166	67	47
201	13	112	147	112	67	67	166	182

TABLE II. Comparison Between Shapley and Centrality Values Using Les Misérables Dataset

DC	BC	CC	EVC	Proposed Model				
				Game 1	Game 2	Game 3	Game 4	Game 5
11	11	11	11	11	11	11	11	11
49	1	56	49	2	55	56	2	56
56	49	28	56	49	49	28	49	26
28	56	26	59	28	43	26	28	49
26	24	49	65	24	44	49	56	28
24	26	59	63	56	73	27	26	27
59	28	27	28	26	32	25	24	70
63	52	65	26	52	51	59	27	69
65	59	69	66	27	57	65	25	71
64	17	70	66	25	40	69	52	42

Similarly, we referred to the work of Hardin, Sarkis and Urc [26] to compare the efficiency of our model using Enron email dataset. Table III shows the results obtained for the same.

We observed that Philip K. Allen, the Managing Director of Trading, appeared in the results of all the coalition games. Mike Grigbsy, VP of Trading, is also an important figure who is present in the results of three of the five games. Found in results of four coalition games, Barry Tycholiz is also the VP of Trading. Another person who can be identified as a prominent figure is Director for State Government, Jeff Dasovich. Game 5 recognizes Louise Kichen – the president of Enron – as one of the most significant nodes.

To get a better numerical understanding of our results, we used a comparison metric – *The Jaccard* Index, also known as the Union Intersection and the *Jaccard* Similarity Coefficient – which is used to calculate the similarity and diversity of sample sets.

TABLE III. COMPARISON OF MOST IMPORTANT NODES USING ENRON DATASET BASED ON VARIOUS CENTRALITY MEASURES

DC	BC	EVC	CC	Game 1	Game 2	Game 3	Game 4	Game 5
Jeff Dasovich	Louise Kitchen	Tana Jones	Robert Benson	Scott Neal	Phillip K. Allen	Kevin Presto	Phillip K. Allen	Scott Neal
Mike Grigsby	Mike Grigsby	Sara Shackleton	Mike Grigsby	Phillip K. Allen	Scott Neal	James D. Steffes	Scott Neal	Mike Grisby
Tana Jones	Susan Scott	Stephanie Panus	Louise Kitchen	Mike Grisby	Mike Grisby	Phillip K. Allen	Mike Grisby	John Arnold
Sara Shackleton	Jeff Dasovich	Marie Heard	Kevin M. Presto	Barry Tycholiz	Barry Tycholiz	Mark Haedick	Barry Tycholiz	John Lavorato
Richard Shapiro	Mary Hain	Susan Bailey	Susan Scott	Sally Beck	Sally Beck	Steven J. Kean	Sally Beck	Joe Quenet
Steven J. Kean	Sally Beck	Kay Mann	Scott Neal	John Lavorato	John Lavorato	Mike Swerzbin	John Lavorato	Phillip Allen
Louise Kitchen	Kenneth Lay	Louise Kitchen	Barry Tycholiz	Susan Scott	Mark Haedick	Jeff Dasovich	Mark Haedick	Barry Tycholiz
Susan Scott	Scott Neal	Elizabeth Sager	Greg Whalley	Kim Ward	Susan Scott	Richard Sanders	Richard Sanders	Sally Beck
Michelle Lokay	Kate Symes	Jason Williams	Phillip K. Allen	Mark Haedick	Richard Sanders	Doug Gilbert-Smith	Kim Ward	Louise Kitchen
Chris Germany	Cara Semperger	Jeff Dasovich	Jeff Dasovich	Bill Williams	Kim Ward	Richard Shapiro	Kevin Presto	David Delainey

The coefficient of Jaccard measures similarity between finite sample sets and is defined as the intersection size divided by the size of the union of sample sets which is shown in (2).

$$i(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (2)$$

We compared the intersection similarity of the most significant nodes from each coalition game, with the results of the proposed model. Finally, for holistic comparison, we took the mean overall intersections, as shown in (3).

$$I_{centrality} = \frac{i_1 + i_2 + i_3 + i_4 + i_5}{5} \quad (3)$$

$I_{centrality}$ depicts the mean of all intersections between sets over the five coalition games, where *centrality* denotes the centrality model used, and *ij* represents the intersection similarity between centrality measures with game j. The results are displayed in Table IV.

TABLE IV. JACCARD SIMILARITY COEFFICIENT VALUES

MEASURE	INTER-SECTION OF SETS		
	USAIR97	ENRON EMAIL	LES MISÉRABLES
I_{dc}	0.762	0.112	0.368
I_{ec}	0.685	0.010	0.354
I_{bc}	0.475	0.156	0.389
I_{cc}	0.514	0.304	0.490

Fig. 1 shows the comparison of the *Jaccard* Indices measured with various degree centralities. I_{dc}, I_{ec}, I_{bc} and I_{cc} denote Degree Centrality, Eigen-vector Centrality, Betweenness Centrality and Closeness Centrality, respectively. USAir97 shows the maximum similarity with I_{dc}, whereas Enron Email and Les Misérables dataset show maximum similarity with Closeness Centrality. Thus we were able to show the comparison of the Shapely algorithm considering EC, DC, BC and CC as benchmarks for all three datasets.

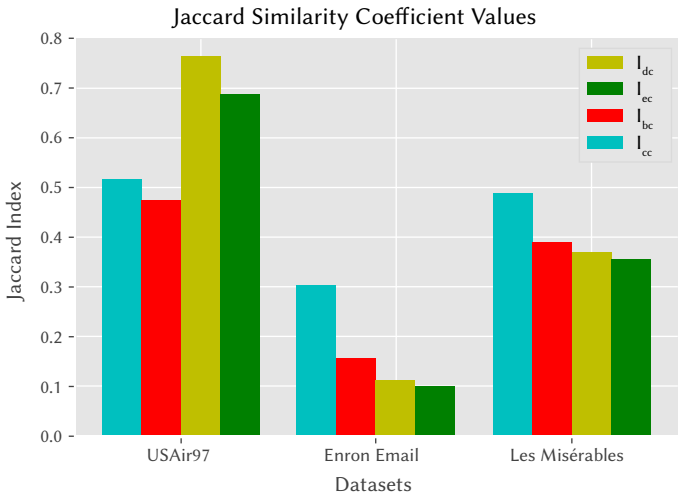

Fig. 1. Comparison of *Jaccard* Indices.

V. DISCUSSION

We had observed many disadvantages in primitive centrality measures that had been used in the past for finding the most influential node, including putting too much focus on the individual node and not on the neighbours of the node. An elaborate description of these disadvantages is mentioned in section II. Game-theoretic approaches like the SV algorithm, take into consideration the marginal contribution of a node to every coalition that it is a part of. This approach has also not been specifically used in the past for RSD problem. For this reason, we aimed to explore the effectiveness of this approach for the purpose of RSD. Our results show a good similarity score (*Jaccard* Index) with the previous studies that used primitive centrality measures.

But as discussed, there were numerous disadvantages with these measures that our SV-based approach tried to overcome. Hence, we observe a slight difference between the most influential nodes found by our approach and those found by the earlier studies conducted on the same datasets.

VI. Conclusion

Sometimes the propagation of rumours on online social networks can lead to serious social problems. It is known to be of great value to accurately identify them from regular comments. Social media rumours have recently become a major concern, especially as people are aware of their ability to influence society. Rumours can not only cause social hysteria in all sorts of crises, but can also cause mass events that are unpredictable and threaten social stability.

We tried to introduce a game-theoretical algorithm in our research work in order to detect the origin of rumour in a complex network. The algorithm used is the algorithm of Shapley. We compared the performance of our game-theoretic approach with prime centrality measures. We also sought to locate prominent top nodes to catch and record multiple potential gossip sources, rather than concentrating discreetly on a single source. The most influential node identified is assumed to be the rumour source in the network.

To evaluate our algorithm on various real-world scenarios, we examined five different game situations, thereby taking into consideration various approaches to determine the most influential nodes in a given dataset. This helped us to gain a deeper and holistic understanding of the game-theoretical algorithm. The *Jaccard* Index has been used as a metric of comparison for our proposed method. The model has shown significant success as the most prominent nodes are successfully identified for both the datasets used.

We are currently working on expanding the theory of Shapley algorithm to consider each person's impact in a social network and thus determine the most serious cause of rumours. We plan to extend the idea of finding the most powerful node in social networks to numerous other similar applications for future work, such as the Internet, or urban networks, and involving a given node in disease dynamics. This will help us understand our algorithm's efficiency and accuracy in multiple applications in the real world.

Further, various optimisation techniques on the SV algorithm, for example, Fuzzy Logic will be implemented for mining much larger social networks and to improve accuracy and other relevant metrics of the project. Fuzzy-based implementation will solve various complexities and limitations that we are currently encountering.

Acknowledgement

We are grateful to the Department of Computer Science and Engineering at Delhi Technological University for presenting us with this research opportunity which was essential in enhancing learning and promoting research culture among ourselves.

References

[1] W. Chen, and S.-H. Teng, "Interplay between social influence and network centrality: a comparative study on shapley centrality and single-node-influence centrality," in *WWW '17: Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 967–976.

[2] H. Wei, Z. Pan, G. Hu, L. Zhang, H. Yang, X. Li, and X. Zhou, "Identifying influential nodes based on network representation learning in complex networks," *PLoS ONE*, vol. 13, 2018, doi: 10.1371/journal.pone.0200091.

[3] J. Zhan, S. Gurung, and S. P. K. Parsa, "Identification of top-k nodes in large networks using katz centrality," *Journal of Big Data*, vol. 4, no. 16, 2017, doi: 10.1186/s40537-017-0076-5.

[4] F. A. Rodrigues, "Network centrality: an introduction," *A Mathematical Modeling Approach from Nonlinear Dynamics to Complex Systems*, Nonlinear Systems and Complexity, Springer, vol. 22, pp. 177-196, 2019, doi: 10.1007/978-3-319-78512-7_10.

[5] B. A. Bhuiyan, "An overview of game theory and some applications," *Philosophy and Progress*, vol. 59, no. 1-2, pp. 111–128, 2018, doi: 10.3329/pp.v59i1-2.36683.

[6] N. Yadati, and R. Narayanam, "Game theoretic models for social network analysis," in *Proceedings of the 20th International Conference Companion on World Wide Web – WWW' 11*, 2011, pp. 291-292.

[7] M. T. Irfan, and L. E. Ortiz, "A game theoretic approach to influence in networks," in *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'11)*, AAAI Press, 2011, pp. 688-694.

[8] T. P. Michalak, K. V. Aadithya, B. Ravindran, N. R. Jennings, and P. L. Szczepanski, "Efficient computation of the shapley value for game-theoretic network centrality," *Journal of Artificial Intelligence Research*, vol. 46, no. 1, pp. 607-650, 2013, doi: 10.5555/2512538.2512553.

[9] R. Narayanam, and Y. Narahari, "A shapley value-based approach to discover influential nodes in social networks," in *IEEE Transactions on Automation Science and Engineering*, vol. 8, no. 1, 2011, pp. 130–147, doi: 10.1109/tase.2010.2052042.

[10] R. Narayanam, and Y. Narahari, "Determining the top-k nodes in social networks using the shapley value," *AAMAS*, pp. 1509-1512, 2008, doi: 10.1145/1402821.1402911.

[11] C.-P. Teo, J. Sethuraman, and W.-P. Tan, "Gale-shapley stable marriage problem revisited: strategic issues and applications (extended abstract)," *Integer Programming and Combinatorial Optimization Lecture Notes in Computer Science*, vol. 1610, pp. 429-438, 1999, doi: 10.1007/3-540-48777-8_32.

[12] P. Domingos, and M. Richardson, "Mining the network value of customers," in *KDD '01: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001, pp. 57–66.

[13] S. Gao, J. Ma, Z. Chen, G. Wang, and C. Xing, "Ranking the spreading ability of nodes in complex networks based on local structure," *Physica A: Statistical Mechanics and its Applications*, vol. 403, pp. 130-147, 2014, doi: 10.1016/j.physa.2014.02.032.

[14] K. Stephenson, and M. Zelen, "Rethinking centrality: methods and examples," *Social Networks*, vol. 11, no. 1, pp. 1-37, 1989, doi: 10.1016/0378-8733(89)90016-6.

[15] L. C. Freeman, "Centraity in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215-239, 1978-79, doi: 10.1016/0378-8733(78)90021-7.

[16] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Identification of influential spreaders in online social networks using interaction weighted k-core decomposition method," *Physica A: Statistical Mechanics and its Applications*, vol. 468, pp. 278-288, 2017, doi: 10.1016/j.physa.2016.11.002.

[17] H.-L. Liu, C. Ma, B.-B. Xiang, M. Tang, and H.-F. Zhang, "Identifying multiple influential spreaders based on generalized closeness centrality," *Physica A: Statistical Mechanics and its Applications*, vol. 492, pp. 2237-2248, 2018, doi: 10.1016/j.physa.2017.11.138.

[18] T. Nie, Z. Guo, K. Zhao, and Z.-M. Lu, "Using mapping entropy to identify node centrality in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 453, pp. 290-297, 2016, doi: 10.1016/j.physa.2016.02.009.

[19] J. Shetty, and J. Adibi, "Discovering important nodes through graph entropy: the case of enron email database," in *KDD '05: Proceedings of the 3rd International Workshop on Link Discovery*, 2005, pp. 74–81.

[20] C. W. Tan, P.-D. Yu, L. Zheng, C.-K. Lai, W. Zhang, and H.-L. Fu, "Optimal detection of influential spreaders in online social networks," in *2016 Annual Conference on Information Science and Systems (CISS)*, 2016, pp. 145-150.

[21] Y. Zhou, C. Wu, Q. Zhu, Y. Xiang, and S. Loke, "Rumour source detection in networks based on the seir model," in *IEEE Access*, vol. 7, 2019, pp. 45240-45258.

[22] F. Liu, and M. Li, "A game theory-based network spreading model: based on game experiments," *International Journal of Machine Learning and Cybernetics*, vol. 10, pp. 1449–1457, 2019, doi: 10.1007/s13042-018-0826-5.

[23] T.-H. Fan, and I.-H. Wang, "Rumor source detection: a probabilistic perspective," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4159-4163.

[24] Y. Zhang, and Y. Zhang, "Top-k influential nodes in social networks: a game perspective," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*, Association for Computing Machinery, New York, NY, USA, 2017, pp. 1029–1032, doi: 10.1145/3077136.3080709.

[25] T. Qiao, W. Shan, and C. Zhou, "How to identify the most powerful node in complex networks? a novel entropy centrality approach," *Entropy*, vol. 19, pp. 614, 2017, doi: 10.3390/e19110614.

[26] J. Hardin, G. Sarkis, and P. C. Urc, "Network analysis with the enron email corpus," *Journal of Statistics Education*, vol. 23, 2015, doi: 10.1080/10691898.2015.11889734.

[27] P. Munjal, N. Arora, and H. Banati, "Dynamics of online social network based on parametric variation of relationship," in *2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, Kolkata, 2016, pp. 241-246.

[28] Bureau of Transportation Statistics. Accessed: July 10, 2020. [Online]. Available: https://transtats.bts.gov/

[29] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, "Simplicial closure and higher-order link prediction," *PNAS*, vol. 115, no. 48, pp. E11221-E11230, 2018, doi: 10.1073/pnas.1800683115.

[30] email-Enron Dataset. Accessed: Jan. 7, 2020. [Online]. Available: https://www.cs.cornell.edu/~arb/data/email-Enron/index.html

[31] Donald E. Knuth, "The stanford graphbase: a platform for combinatorial computing," Association for Computing Machinery, New York, NY, USA, 1993.

[32] H. Yang, J. Luo, Y. Liu, M. Yin, and D. Cao, "Discovering important nodes through comprehensive assessment theory on enron email database," in *3rd International Conference on Biomedical Engineering and Informatics*, Yantai, 2010, pp. 3041-3045.

Minni Jain

Minni Jain is currently working as an Assistant Professor in Delhi Technological University, Delhi, India. She obtained her M. Tech. degree in Information Security and B. Tech. degree in Information Technology. Her major research interests include Natural Language Processing, Sentiment Analysis, Information Security, Neural Networks and Fuzzy Logic.

Aman Jaswani

Aman Jaswani obtained his B. Tech. degree in Software Engineering from Delhi Technological University, Delhi, India. He has worked on numerous projects in the field of Machine Learning and Natural Language Processing. He plans to pursue a Masters Degree in a similar field. His areas of research include Game Theory, Machine Learning, Natural Language Processing and Data Science.

Ankita Mehra

Ankita Mehra obtained her B. Tech. degree in Software Engineering from Delhi Technological University, Delhi, India. She is a final year student. Her areas of research include Game Theory, Machine Learning, Deep Learning, Computer Vision and Natural Language Processing.

Laqshay Mudgal

Laqshay Mudgal obtained his B. Tech. degree in Software Engineering from Delhi Technological University, Delhi, India. His major research interests include Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, and Data Science.