Research Article **Open Access**

Xue-Guang Wang*

Research on Critical Nodes Algorithm in Social Complex Networks

DOI 10.1515/phys-2017-0008
Received Aug 10, 2016; accepted Sep 14, 2016

Abstract: Discovering critical nodes in social networks has many important applications and has attracted more and more institutions and scholars. How to determine the K critical nodes with the most influence in a social network is a NP (define) problem. Considering the widespread community structure, this paper presents an algorithm for discovering critical nodes based on two information diffusion models and obtains each node's marginal contribution by using a Monte-Carlo method in social networks. The solution of the critical nodes problem is the K nodes with the highest marginal contributions. The feasibility and effectiveness of our method have been verified on two synthetic datasets and four real datasets.

Keywords: Critical Node Problem; Complex Networks; Community Structure; Influence Maximization

PACS: 01.20.+x; 07.05.Mh

1 Introduction

The analysis of social networks has many important applications and has attracted more and more institutions and scholars. How to identify influential individuals from a network is called a critical node problem (CNP), and this is one of the basic problems in the field of social network analysis. An effective solution for this problem will have an important practical value [1]. For example, we can better target and isolate the source of the disease and block its spread and diffusion in the disease network; or companies may identify individuals with influence and let them recommend products to their friends so that a cascade spreads by the greatest extent.

Domingos and Richardson first studied the CNP as an algorithmic problem and presented a greedy approximation algorithm. Currently, the research on the CNP of social networks mainly concentrates on improving the greedy algorithm or developing new methods by graph theory and complex network theory. Considering the widespread community structure in social networks, this paper gives a solution for the CNP, which assigns a marginal contribution for every node in a community of a social network using the solution concept and union concept of cooperative games. Then we sort all nodes by their contribution and obtain critical nodes according to some rules.

2 Background

The models for information propagation in networks have been widely studied [2]. We consider two models in this paper: independent cascade model (ICM) and linear threshold model (LTM) [2].

1) ICM Model. In this model, a propagation probability $p_{u,v}$ is given for each edge $(u, v) \in E$, that is, vertex v is activated with probability $p_{u,v}$ by u. When an initial set A_0 of active vertices is given, the diffusion process spreads up according to the following randomized rule. When a vertex u is activated at time-step i, it has a single chance of activating its neighbor v with $p_{u,v}$. If u succeeds, v will become active at time-step $i + 1$. Here, if v has multiple parent vertices that become active at time-step i for the first time, then their activation attempts are sequenced in an arbitrary order. Whether or not u succeeds, it cannot make any further attempts to activate v in subsequent steps. The process runs until no more activation is possible.

2) LTM Model. In this model, vertex v is influenced by each neighbor u according to a weight $w_{v,u}$ such that $\sum_v w_{v,u} \leq 1$. Each vertex has a predefined threshold $\theta_v \in [0, 1]$, which is chosen uniformly at random. When an initial set A_0 with active vertices is given, the diffusion process unfolds according to the following randomized rule. All activated vertices at time-step i still keep active at time-

***Corresponding Author: Xue-Guang Wang:** Department of Computer Science, East China University of Political Science and Law, Shanghai 201620, China, Department of Computer Science, University College London (UCL), LondonWC1E 6BT, United Kingdom of Great Britain and Northern Ireland;
Email: wangxueguang@ecupl.edu.cn

© 2017 X.-G. Wang, published by De Gruyter Open.
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License.

Unauthenticated
Download Date | 11/13/19 4:40 PM

step $i+1$. Whether or not any inactivated vertex is activated is determined by its neighbors' weights such that $\sum_v w_{v,u} \geq \theta_v$. The process runs until no more activation is possible.

The difference between ICM and LTM is that (for ICM,?) each attempt of activation is independent of the attempts by all the other active individuals, while in the latter model each inactive individual is influenced by the aggregated weight of all its active neighbors.

A network is modeled as a graph $G = (V, E)$ with vertices in V modeling the individuals in the network and edges in E modeling the relationship between individuals. A vertex has two states: active and inactive, which in this context means whether a product or idea is accepted by individuals or not. We assume that a vertex can only change from inactive to active and not vice versa; an inactive vertex can be activated by its active neighbor vertices and an active vertex can activate its inactive neighbors; the increment of activated vertices represents the dissemination of information.

The definition of the CNP is as follow:

Input: $G = (V, E)$ and an integer k;

Output: $A = \arg\max_{S \subseteq V, |S| \leq k} \sigma(S)$

For $k - CNP$, we hope to find out a set A with k elements and maximize $\sigma(A)$ which is a NP-hard (define). But, Nemhauser et al. have approved that there is a greedy algorithm (see Algorithm 1) which approximates the optimum within a factor of $(1 - 1/e)$[4].

Algorhm 1 Greedy Algorithm

1: $A = \Phi$
2: **for** $i = 1$ to k
3: $v_i = \arg\max_{u \in V \setminus A}(\sigma(A \cup \{u\}) - \sigma(A))$
4: $A = A \cup \{v_i\}$
5: **end for**

There is a key problem in how to compute the value of $\sigma(A)$ in Algorithm 1. Currently, we do not have any efficient method to obtain its exact solution. However, we can use a Monte-Carlo method to simulate the process of influence spread for obtaining approximate results by high probability.

3 The Algorithm

3.1 Introduction to the algorithm

Given a finite set of players N, a cooperative game with transferable utility as a pair (N, v), characteristic function $v : 2^N \to \mathrm{R}$ and $v(\emptyset) = 0$. For $\forall i \in N$, if payoff vector satisfies $x_i \geq v(\{i\})$ and $\sum_{i=1}^{N} x_i = v(N)$, then it is so-called an allocation of (N, v). The solution of a cooperative game is a kind of allocation rule and the allocated payoff for every player denotes a method to measure the negotiation strength of the players in the game. Shapely (cite) presented a solution concept which determines the only allocation distribution scheme from the solutions with different property, i.e., it assigns the player's payoff according to the importance of every player for the game [5]. The Shapely value of the player i in the game (N, v) is

$$Sh_i(v) = \sum_{\{S \subseteq N | i \in S\}} \frac{(n-s)!(s-1)!}{n!}(v(S \cup \{i\}) - v(S)), \quad \forall i \in N$$

where $n = |N|$ and $s = |S|$.

However, the Shapely value does not consider the impact of coalition structure and Owen extends it [6]. Each union obtains its payoff from the game between the unions, and then the payoff is allocated by the internal game among the members of the union. All the payoffs are computed by the Shapely value.

Assuming that $N = \{1, 2, \ldots, n\}$ and $M = \{1, \ldots, m\}$, a partition $P = \{N_1, N_2, \ldots, N_m\}$ is a coalition structure on N. Let N_k a union and $\bigcup_{1 \leq k \leq m} N_k = N$. When $l \neq k$, $N_l \cap N_k = \emptyset$. For $i \in N$, $k(i)$ denotes the index of the union containing player i, so $k(i)$ is defined by the relation $i \in N_{k(i)}$. For $k \in M$ and $S \subseteq N_k$, the game $\hat{v}_S$ is defined by

$$\hat{v}_S(Q) = \begin{cases} v(\cup_{h \in Q} N_h), & k \notin Q \\ v(\cup_{h \in Q \setminus \{k\}} N_h \cup S), & k \in Q \end{cases}$$

where $Q \subseteq M$.

The game $\bar{v}_k$ is defined by $\bar{v}_k(S) = Sh_k(\hat{v}_S)$, then the Owen value of the player $i \in N$ in the game $\bar{v}_{k(i)}$ is $Ow_i(v, P) = Sh_i(\bar{v}_{k(i)})$.

3.2 A new algorithm

Because the community structure is prevalent in social networks [7], we respectively consider the community's influence on the information diffusion and every node's influence in the community. We take the nodes in the social network as the players in the cooperative game and information diffusion as coalition formulation. We can map the information diffusion into a cooperative game with transferable utility which is formalized in the literature [8]. So, we can use the Owen value to obtain the marginal contribution of every node. Because the Owen value can be seen as a two-step procedure in which the Shapely value applies twice, we firstly compute a node's Shapely value.

Unauthenticated
Download Date | 11/13/19 4:40 PM

Given a node $i \in N$ and a subset $S \subseteq N$ such that $i \notin S$, the marginal contribution of node i is $v(S \cup \{i\}) - v(S)$, $\forall S \subseteq N \backslash \{i\}$. Consider the set of all possible permutations Ψ on N, let $\psi \in \Psi$ and define $S_i(\psi)$ to be the set of all nodes appearing before node i in the permutation ψ. So, the average marginal contribution of node i to the given coalition game is

$$\frac{1}{n!}\sum_{\psi\in\Psi}[v(S_i(\psi)\cup\{i\}) - v(S_i(\psi))].$$

Note that, this method must work with $n!$ permutations and its computational complexity is $O((n/e)^n)$[9]. Therefore, we give the approximate method for computing the Shapely value. Randomly generated t-sets Ψ_t with t permutations, see Algorithm 2, let $\psi \in \Psi_t$ and $\psi(i)$ denotes the ith node in the permutation. The number of activated nodes after running the diffusion model when the node $\psi(1)$ is activated is the contribution of $\psi(1)$. Next, we consider the node $\psi(2)$. If $\psi(2)$ becomes active after $\psi(1)$ is activated, then the contribution of $\psi(2)$ is 0. Otherwise, the contribution of $\psi(2)$ is the number of activated nodes by $\psi(2)$. Therefore, we can obtain the contributions of $\psi(3),\ldots,\psi(n)$. For $\psi \in \Psi_t$, we repeat the above process R times. Then the average contribution of each node in the diffusion process can be calculated. We can obtain the top-k nodes by sorting by the greatest influence and ensure that they are not adjacent to each other. See Algorithm 3 and Algorithm 4.

Algorithm 2. CREAT_Ψ_t(n)

```
t=random(n)
m=random(n)
Select_t (n)
for i = 0 to m do
   arr[i] = random() % (n-i)
  for j=0 to i do
    if arr[j] <= arr[i] then
      arr[i]=arr[i]+1
    end if
  end for
  v = arr[i]
  for k = i-1 to j do
    arr[k+1] = arr[k]
    k=k-1
  end for
  arr[j] = v
end for
for h=2 to t do
   |Ψ_{h-1}| =arr
   n=n-m
   m=random(n)
   Select_t(n)
end for
|Ψ_h| =arr
```

Algorithm 3. ShapelyValue(v, R)

```
n=|ψ|
t=|CREAT_Ψ_t|
tmp[1...n]=0
ShV[1...n]=0
for i=1 to t do
  for r=1 to R do
    for j=1 to n do
      tmp_b=v(S_j(ψ_i) ∪ {j}) − v(S_j(ψ_i))
      tmp[j]=αtmp[j] + (1-α)tmp_b
    end for
  end for
end for
for i=1 to n do
  ShV[i]=tmp[i]/t
end for
```

Algorithm 4. TOP_K(V)

```
TopK[1...k]=0
AsceSort(V);
TopK[1]= V[1]
i=1
j=2;
while i < k do
  if V[j] is not adjacent to TopK[1...i-1] then
    TopK[i]= V[j]
    i=i+1
  end if
  j=j+1
end while
```

According to the description in Section 3.1, we provide the calculation method of the Owen value. Roger Guimerà, et al studied node roles in a community according to within-module degree z and participation coefficient P, which are divided into seven categories $Role = \{R1, R2, R3, R4, R5, R6, R7\}$[10]. We consider two types of the roles: Non-hub connector node ($z < 2.5$ and $0.62 < P \leq 0.80$, R3) and Connector hub ($z \geq 2.5$ and $0.30 < P \leq 0.75$, R6). Most of the nodes which belong to these two roles connect to other communities.

We use the CNM algorithm [11] to divide the network $G = (V, E)$ into l communities $C = \{C_1, C_2, \ldots, C_l | C_i = (V_i, E_i),\ i = 1, \ldots, l\}$ and assign the role $rv_i^h \in Role$ for every node $v_i^h \in V_i$ in the community. Let $Role' = \{R3, R6\}$, $i = 1, \ldots, l$, $j = 1, \ldots, l$,

$$V' = \{v_i^h, v_j^g | \exists (v_i^h, v_j^g), v_i^h \in V_i, rv_i^h \in Role', i \neq j\},$$

Unauthenticated
Download Date | 11/13/19 4:40 PM

$$E' = \{e_i^h, (v_i^h, v_j^g) | e_i^h = (v_i^{h_a}, v_i^{h_b}) \in E_i, v_i^{h_a} \in V_i, v_i^{h_b} \in V_i, rv_i^h \in Role', i \neq j\},$$

define $G' = (V', E')$ as the community game network. So, we can obtain the Shapely value of every node in the network G' and take the sum of the Shapely values of all nodes in the same community as the community's payoff. Then we treat a community as a separate network and calculate the Shapely value of every node in the community. The node's Owen value is assigned according to the normalized Shapely value of the node in the community and the community's payoff. So, we can obtain the *k*critical nodes by Algorithm 4.

4 Experiments

We validate our method using two synthetic datasets and four real network datasets. All experiments are executed using a PC with 3.2GHz CPU, 4G Memory and Windows 7. The development tools are Matlab 2009 and Microsoft Visual Studio 2010.

4.1 Datasets

We use the BA model [12] and the Forest Fire model [13] to generate two synthetic datasets with 5000 nodes. The BA model takes the power-law distribution and has two important features: growth and preferential attachment. The model can generate a scale-free network whose power exponent is 3 and community structure is not obvious. The Forest Fire model can generate a network with degree power law, densification law, shrinking diameter, and obvious community structure.

The real datasets include DBLP, Facebook, Enron and Youtube. The DBLP dataset [14] constructs a co-author network with 143276 nodes and 359812 edges according to the papers published in the conferences and journals of the computer field from 1997 to 2006. The Facebook dataset [15] provides the friendship network of the New Orleans area with 60567 users and 583766 connections obtained from Jan. 1, 2007 to Dec. 31, 2008. The Enron dataset [16–19] is an E-mail network with 36692 nodes and 367662 edges. The Youtube dataset [20, 21] consists of 35468 nodes and 261191 edges obtained from Jan. 1, 2007 and Jan. 15, 2007.

4.2 The results

We compare our method CNP-OV with the greedy algorithm GREEDY (Algorithm 1), the degree-heuristic algorithm DEGREE and the random algorithm RANDOM. DEGREE selects k nodes with the greatest degree from the network as the initial set; RANDOM randomly selects k nodes as the initial set. In order to obtain the accurate influence of every algorithm, we use the average number of the activated nodes after running ICM and LTM 10000 times for every initial set. In ICM, the propagation probability is set to 0.05; in LTM, the edge-weight of a node is the reciprocal of the node's degree. The size of the initial set is respectively set from 1 to 20. In the experiments, we only discuss the case of using the ICM because we obtain the same conclusions for ICM and LTM.

First, we consider the influence of community structure on the CNP-OV. We respectively compute the Shapely value and Owen value of every node in the BA and FF datasets and obtain the initial set and the number of activated nodes after running the ICM. The process is repeated 100 times and the average number of the nodes activated

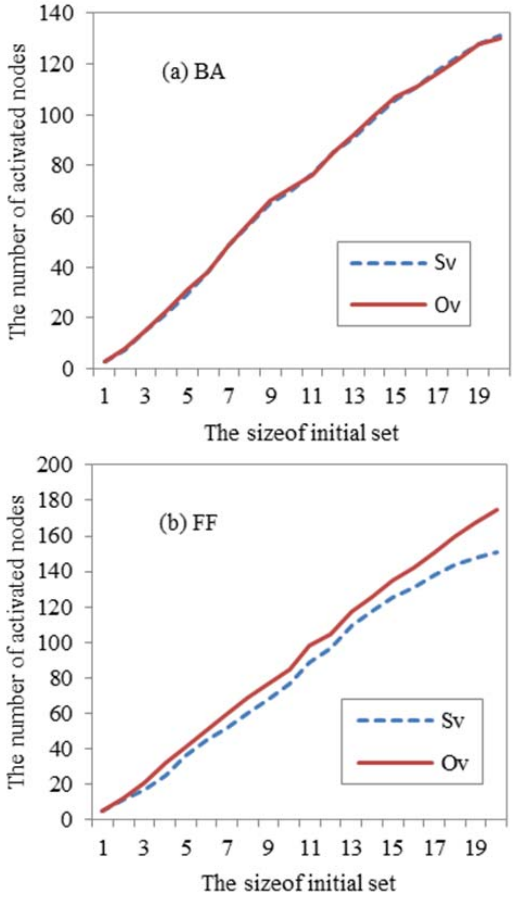

Figure 1: The influence of community structure on our method.

Unauthenticated
Download Date | 11/13/19 4:40 PM

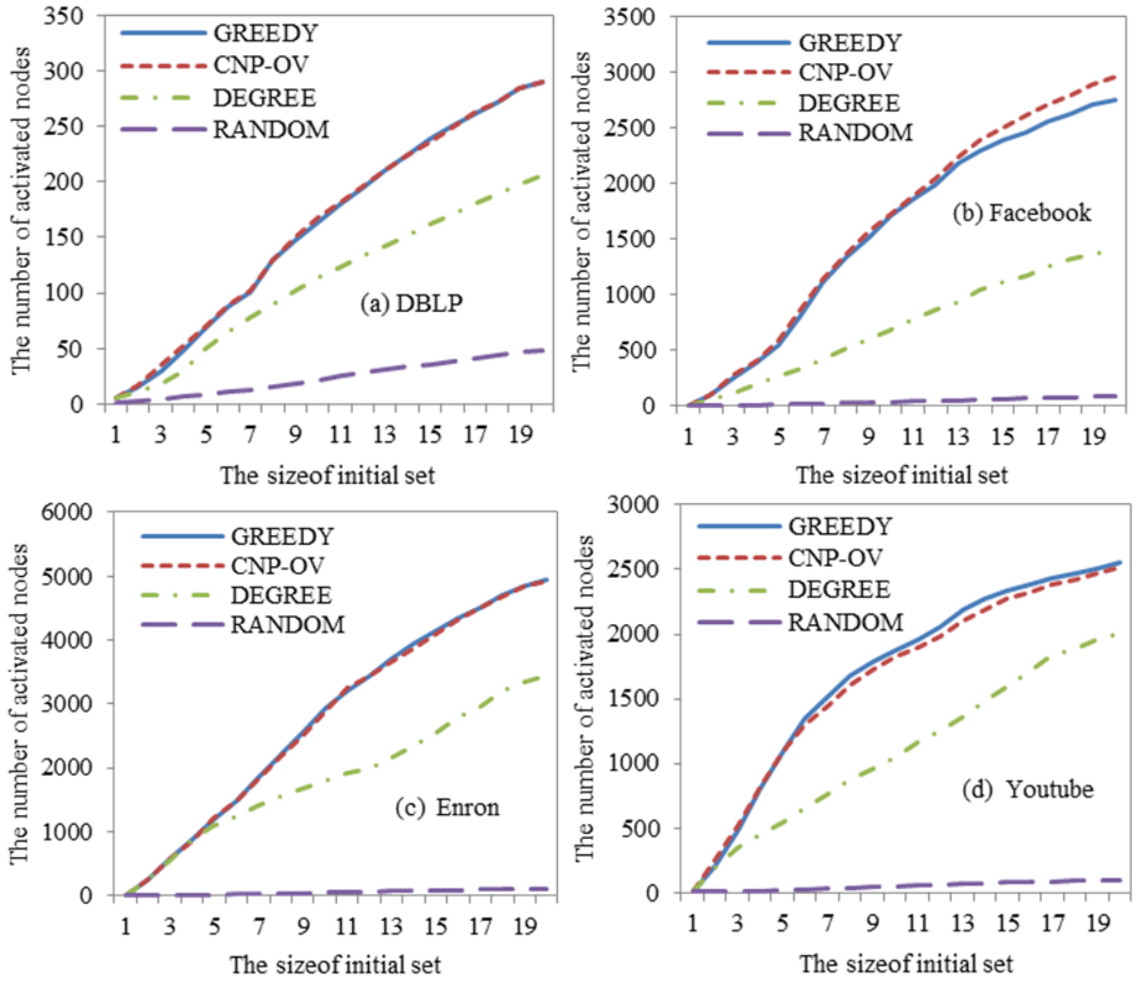

Figure 2: Performance Analysis

by the initial set with different size is drawn in Figure 1. The dashed Sv denotes the method based on the Shapely value and the solid line Ov is the method based on the Owen value. The Sv is almost the same as the Ov for the BA dataset with an unobvious community structure (Figure 1a). In contrast, the Ov is significantly better than the Sv for the FF dataset with an obvious community structure (Figure 1b).

Secondly, we analyze our method's accuracy. We respectively use GREEDY, DEGREE, RANDOM and CNP-OV to determine the initial sets from above for four real datasets. Figure 2 describes the number of nodes activated by the initial set with different size based on ICM. The accuracy of the CNP-OV is similar to GREEDY, sometimes even better than it (Figure 2b). Compared with DEGREE and RANDOM, the CNP-OV provides a large advantage.

5 Conclusions

Considering the widespread community structure in social networks, for solving the CNP, in this paper we present a new algorithm for discovering critical nodes based on cooperative games. It assigns a marginal contribution for every node in a community of social networks using the solution concept and union concept of cooperative games. We sort all nodes by their contribution and obtain critical nodes according to selected rules. We have validated its feasibility and effectiveness. The next step will be to study how to improve the time efficiency of the CNP-OV.

Acknowledgement: This work is supported by the National Social Science Foundation of China (No.11BFX125), PuJiang Talent Project, Peak of law subject construction Project and Public Security Discipline Construction Foundation.

Unauthenticated
Download Date | 11/13/19 4:40 PM

References

[1] He N., Li D.Y., Gan W.Y., Zhu X., Mining vital nodes in complex networks, Computer Science, 2007, 34(12): 1-5.

[2] Even-Dar E., Shapira A., A note on maximizing the spread of influence in social networks, WINE 2007, LNCS 4858, 2007: 281-286.

[3] Shetty J., Adibi J., Discovering Important Nodes through Graph Entropy The Case of Enron Email Database, The 3rd international workshop on Link discovery, Chicago, Illinois, 2005:74-81.

[4] Nemhauser G., Wolsey L., Fisher M., An analysis of the approximations for maximizing submodular set functions, Mathematical Programming, 1978, 14(1): 265-294.

[5] Shapley L.S., A value for n-person games, In: Kuhn H W, Tucker A W (Eds.), Contributions to the Theory of Games II, Princeton University Press, 1953, 307-317.

[6] Owen G., Values of games with a priori unions, In: Henn R, Moeschlin O (Eds.), Essays in mathematical economics and game theory, Springer-Verlag, Berlin, 1977, 76-88.

[7] Scott J., Social Network Analysis: A Handbook (2nd ed), Sage, London, 2000.

[8] Jean-Francois C., Network Games as TU Cooperative Games: The Core, the Shapley Value and Simple Network Games, 2009, http://centres.fusl.ac.be/CEREC/document/seminars/caulier_cerec_feb2009.pdf

[9] Cormen T.H., Leiserson C.E., Rivest R.L., Stein C., Introduction to Algorithms, 2nd ed, Cambridge, MA: MIT Press, 2001.

[10] Guimerà R., Amaral L.A.N., Functional cartography of complex metabolic networks, Nature, 2005, 433:895-900.

[11] Clauset A., Newman M.E.J., Moore C., Finding community structure in very large networks, PHYSICAL REVIEW E, 2004, 70(6): 066111 (6).

[12] Barabási A.L., Albert R., Emergence of scaling in random networks, Science, 1999, 286(5439): 509-512.

[13] Leskovec J., Kleinberg J., Faloutsos C., Graphs over time: Densification laws, shrinking diameters and possible explanations, The 11th ACM SIGKDD international conference on Knowledge discovery in data mining (KDD). Chicago, Illinois, USA, 2005: 177-187.

[14] http://dblp.uni-trier.de/

[15] Viswanath B., Mislove A., Cha M., Gummadi K.P., On the evolution of user interaction in Facebook, The 2nd ACM SIGCOMM Workshop on Social Networks (WOSN), Barcelona, Spain, 2009: 37-42.

[16] http://www-2.cs.cmu.edu/~enron/

[17] Awadalla N.S., Hanna M.A., Ismail M.N., Period Variation Study and Light Curve Analysis of the Eclipsing Binary GSC 02013-00288, Applied Mathematics and Nonlinear Sciences, 2016, 1(2): 321–334.

[18] Chen Q., Chang H., Govindan R., Jamin S., Shenker S.J., Willinger W., The origin of power laws in Internet topologies revisited, The 21st Annual Joint Conference of the IEEE Computer and Communications Societies, Los Alamitos, CA, USA, 2002: 608-617.

[19] Watts D.J., Strogatz S.H., Collective dynamics of “small-world” networks, Nature, 1998, 393: 440-442.

[20] Rosa M., Gandarias M.L., Multiplier method and exact solutions for a density dependent reaction-diffusion equation, Applied Mathematics and Nonlinear Sciences, 2016, 1(2):311–320.

[21] Mislove A., Marcon M., Gummadi K.P., Druschel P., Bhattacharjee B., Measurement and analysis of online social networks, The 7th ACM SIGCOMM conference on Internet measurement conference (IMC), San Diego, California, USA, 2007: 29-42.

Unauthenticated
Download Date | 11/13/19 4:40 PM