Discovering Important Nodes in Social Networks Using Entropy Measure

Vahid Bashiri, Hossein Rahmani*, Hamid Bashiri
Iran University of Science and Technology
Tehran, Iran
vbashiri@comp.iust.ac.ir, h_rahmani@iust.ac.ir, hbashiri@comp.iust.ac.ir

Received: 2017/06/20 Revised: 2017/07/01 Accepted: 2018/03/14

***Abstract*— Discovering important nodes in graph data attracted a lot of attention. Social networks are good examples of graph data in which each node represents a person and each edge represents a relationship between two people. There are several methods for the task of discovering important nodes in graph data. In this paper, important people are defined with their roles in society or organization. We propose an efficient method to discover leaders in graph network. For this purpose, both structural feature like entropy and inherent features including from, to, subject and message's time of social networks are used to propose a novel method for discovering important nodes in social networks. The proposed method was applied to Enron dataset and compared with previous methods. The proposed method succeeded to first, discover more important roles in Enron dataset, second, determine CEO as leader of Enron Corporation and third, discover two out of four CEOs among top VIPs.**

Keywords— Graph Mining; Social Network; Important Node; Entropy;

1. INTRODUCTION

Finding important nodes in graph data attracted a lot of attention. This issue has been considered in many domains such as economy [1] , biology [2, 3], chemistry [4], and others [5, 6]. As a good sample of graph data, we could mention to social networks. By increasing the use of social networks among the majority of people, considered as a good representation of real society therefore, automatic analysis of people's behavior becomes possible. Among many analyses, discovering important persons in social networks had many advantages [7], yet it encountered many challenges. In the graph model of real society, each node represents a person and each edge indicates kind of relationships between two connected persons [8, 9, 10]. In the real society, types of relationships varied from kinship (i.e., father, mother, brother, sister and so on), workplace relationship (i.e., lower ranks to superior, chairman to CEO, etc.) and friendship. By discovering important nodes in social networks it is possible to: Firstly, detecting the most influential people in social networks which could be used as staring points in viral marketing [11]. Secondly, analyze the interaction between important persons and ordinary people in the society and eventually, predict/maintain/manage the crowd's reactions according to important person's interactions [12]. Accordingly, the whole network could be controlled by controlling small set of important nodes.

Emails are good indicator of collaborations between two persons. Email logs can be used to create social network [13, 14, 15]. Two or more persons could communicate easily through emails. In addition to the ease of use, emails are robust (comparing to mobile and telephone), easy future retrieval (your communication content is available in future for retrieval and references) and asynchronous (there is no need that communicative people present in the same time in the communication channels). So, emails are considered as one of the main communication channels in many organizations. Additionally, from the analytical point of view, existence of additional features such as 1-Sent time, 2-Subject, 3-Contact person and 4-Content, makes emails as a good sample of social network for discovering important nodes in graph data.

Whereas some of the previous methods using standard graph measures such as closeness centrality, betweenness centrality and so on to discover important nodes in graph data, in this paper, it is assumed that important people discuss more varied set of subjects. Accordingly, each person interaction (consider the both subjects and the interacted persons at the same time) was presented as graph; then we introduce new way of calculating entropy for each node in graph data. Finally, email's specific features e.g., time of sent mails and number of sent and received mails were considered for the task of discovering important nodes. The variables introduced in *Table 1*.

The rest of this paper is organized as follows: Section 2 discusses the previous work in the field of discovering important nodes. In Section 3, we discuss our proposed method. In Section 4, we evaluate our method and compare our proposed method with previous methods. In Section 5, we conclude and discuss future trends for our proposed method.

1. PREVIOUS WORK

Kooti et al [16] study 16 billion emails exchanged between 2 million yahoo accounts to predict replying behavior of users. Alsmadi et al [17] using his personal email dataset try toperform clustering and classification. Getoor et al [18] use node similarity and clustering approach to discover important nodes. Freeman et al [19] use centrality measure for this purpose. Kaur et al [20] use

TABLE I. PARAMETERS AND THEIR DISTINCT VALUES

Variable	Distinct Values	Variable Type	Description
Sent_{count}	0 -∞	Numerical	shows the number of mails sent
Receive_{count}	0 -∞	Numerical	shows the number of mails received
Working_{hour}	{0,1}	Boolean	shows an email sent in or out of work hour (between 8 to 17) 0 = out of work time 1 = in work time
Status	- CEO - President - Vice President - Managing - Director - Director - in House Lawyer - Manager - Trader - Employee - N/A	Nominal	show position of employees

eigenvector in order to find important nodes. Noble et al [21] use entropy measure to cluster different graphs and accordingly detect anomalies. Kajdanowicz et al [22], use the entropies of centrality measure distributions such as degree centrality, betweenness centrality, closeness centrality to compare real-world graphs with the most prominent graph generation models (Erdős–Renyi random graph model, Watts–Strogatz small world model, Albert–Barabási preferential attachment model, Price citation model). White et al [23] consider several new algorithms such as graph-theoretic notions of weighted paths and Markov chain models to propose a general framework to discover important nodes in graph data. Newman [24] applied standard graph measures such as closeness and betweenness for the task of discovering important nodes. Zhang et al [25] propose a multi-criteria evaluating method based on principal component analysis (PCA) to identify key nodes in graph data. Huang et al [26] proposes an effective ranking method based on degree and betweenness values. Degree Centrality (DC), Betweenness Centrality (BC) and Closeness Centrality (CC) are the methods that are typically used in complex networks [27] to discover important nodes. Wang et al [28] consider the degree of the nodes and degree of their neighborhoods for this purpose. Chen et al [29] propose semi-local centrality measure as a trade-off between the low-relevant/low cost measures such as degree centrality and other more-relevant/time-consuming measures. Saito et al [30] introduce a method in order to find influential nodes in a social network. Also Node importance is highly depends on graph's subject. Xue et al [31] review graph-theoretic node importance mining in world city networks and compare different methods.

2. PROPOSED METHOD

2-1. Graph Construction

In this section, people's email communication modeled as a graph which was inspired by our hypothesis for discovering important nodes in social networks. It is assumed that important nodes interacted with more varied people and discuss diverse subjects. So, variety in both people and subjects should be considered.

For each email c_{ij} with subject s_{ij} between two persons p_i and p_j: At first a new node p_i (if not existed before) is to be added. The same procedure is done for the new node v_j, finally s_{ij} is tokenized into its tokens $w_1 \ldots w_n$ (which are separated by space). For each token w_i that ever appeared in email's subject between p_i and p_j, new directed edge e_{ij}^{w} was added between two persons p_i and p_j. Each e_{ij}^{w} is weighted on a particular account showing the number of times token w appeared in subject part of emails exchanged between p_i and p_j. Now, there is a novel graph representation of Enron dataset in which each node represents an email address and each edge $e_{ij}^{w:n}$ indicate that token w appeared n times in the subject emails of two persons p_i and p_j. Comparing to the previous methods, in our proposed graph representation there could be more than one edge e_{ij} between two persons p_i and p_j.

2-2. Using Entropy to Measure Variety

The proposing hypotheses is that important persons interacting with more distinct people and discuss more widespread subjects. Entropy was used to measure variety which functioned to discover the important person. Although entropy has been precisely defined in computer science [32], there has been several disputes [33] over a consent formula for measuring entropy in graph data [22, 34, 35, 36]. For example, Kolmogorov definition on adjacency matrix is applicable in theory but it is not computable in practice [37]. In this paper, Shannon (1) is used to calculate the entropy in graph data.

$$\mathrm{H}(G) = -\sum_{i=1}^{n} P_i \log P_i \qquad (1)$$

In formula (1), p_i is the probability rate (respective to the set for which entropy is calculated) and n is number of elements in the set. Clearly, sum of the total probability p_i should equals to one., Entropy's value varies from *0* to infinite (more precisely according to (1), *H(G)* varies from *0* to *lnn*).

In order to normalize the entropy values, we could divide entropy to *lnn* (2) which makes the upper-bound of entropy values from *lnn* to *1* [38].

$$\mathrm{H}_o = \frac{H}{\ln n} \qquad (2)$$

In the next step, Formula (1) is used to calculate the score for each node then the entropy value is to be normalized. The probability p_i is calculated by dividing the weight w_i by total weight of connecting edges. *Fig. 1* shows the email communications among three persons p_1, p_2 and p_3. Person p_1 communicates with person p_2 on tokens w_1 and w_2 while simultaneously communicates with person p_3 only on token w_3. According to (2), entropy score p_1 is calculated as follows:

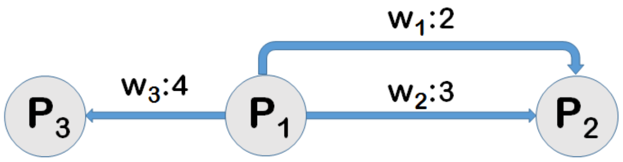

Fig. 1. Entropy example

$$\mathrm{H}(P_1) = -\frac{4}{9}\ln\frac{4}{9} - \frac{2}{9}\ln\frac{2}{9} - \frac{3}{9}\ln\frac{3}{9} = 1.06 \tag{3}$$

According to Formula (2), people with different edge weight who are also in contact with more people get lesser score for their entropy value.

2-3. Final Score

In this section, in addition to entropy measure discussed in Section 3.2, Enron's inherent features were used to augment the process of discovering important nodes. It is assumed that important people have higher ratio of sent e-mails compare to the received ones and also they sent their emails in working hour time (between 8 to 17). According to assumptions above, a formula is proposed that considers the 1-Ratio of number of sent emails divided to sum of sent and received emails (4) and 2-Ratio of number of emails sent in working hours divided to all the emails sent in working and non-working hours (5).

$$\mathrm{S}_o = \frac{Sent_{count}}{Sent_{count} + Receive_{count}} \tag{4}$$

$$\mathrm{T}_o = \frac{WH_{count}}{WH_{count} + NoWH_{count}} \tag{5}$$

In Formula (5), $\mathrm{WH}_{\mathrm{count}}$ and $\mathrm{NoWH}_{\mathrm{count}}$ indicates Working-Hour and Non-Working-hour, respectively.

Formula (6) integrates different measures for discovering important nodes in social networks.

$$\mathrm{Final\_Score} = (1 - \mathrm{H}_0) + \mathrm{S}_0 + \mathrm{T}_0 \tag{6}$$

3. Empirical Results

3-1. Dataset

In Section 1, it is discussed that emails are good representer of people’s social interactions. In this paper, Enron dataset [39] was used because of:

- Enron is the real email dataset that belong to big company and is published online.
- This dataset contains both personal and professional related emails.
- Enron dataset is good source of information for analyzing employee’s interaction in big companies.
- The characteristics of this dataset is similar to the other domains such as fraud and terrorism, so our proposed method could be used in these domains as well.
- One of the major challenges in the problem of important node detection is evaluating the final results. To verify that the discovered important nodes do play a key role in real world, the position of each employee could be used. Access to the recent information on Enron dataset makes the evaluation of final results possible.
- There are several methods that use Enron dataset to evaluate their methods [40, 41, 42, 43]. So, our results could be compared with their results on the same dataset.

Enron dataset was made public by the Federal Energy Regulatory Commission during its investigation. In this paper, the revised version built by Ruhe et al [44] was used. This dataset contains 517,431 emails circulated around 150 employees.

Some additional information such as sender, receiver, date, time, subject and content of emails are also available for analysis. As we mentioned earlier, employee’s position is also available which we use it, mainly, for evaluating our final results.

3-2. Empirical Results

A number of 70000 emails from Enron dataset were selected with uniform sampling. Following the graph construction process discussed in Section 3.1, the graph was built with 9661 nodes and 200597 weighted edges. *Fig. 2* shows the resulted graph using Cytoscape [45]. *Table 2* shows basic statistical information of built graph.

We use formula (2) to calculate the normalized version of entropy and assign to each node an initial score. *Table 3* shows 5 persons with the highest initial scores.

Using Formula (6) improves the result. *Table 4* shows 5 persons with highest scores calculated according to Formula (6). As it is shown in *Fig. 3*, Lavarato has in average 15.25 emails to each president, 8.5 emails to CEOs and 8.9 to vice presidents but only 3.3 emails to all other email addresses.

TABLE II. BASIC STATISTICAL INFORMATION OF BUILT GRAPH.

Number of nodes	7,946
Number of edges	198,883
Average degree	25.0

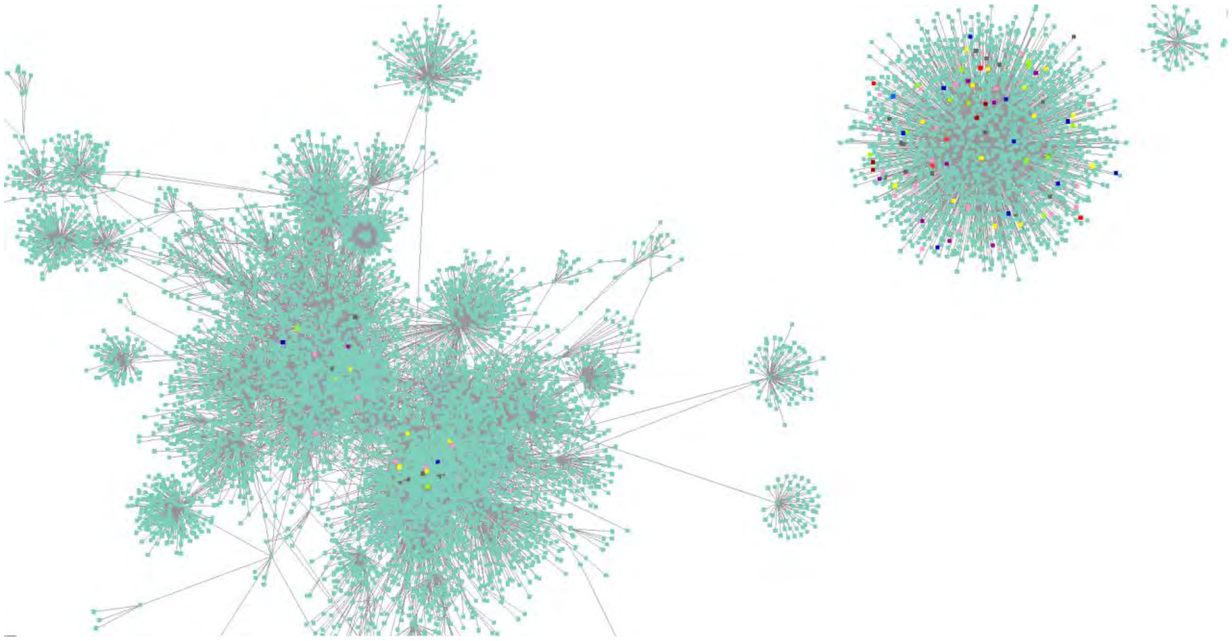

Fig. 2. Enron's Network Graph2: [Brown] CEO, [Red] President, [Yellow] Vice President, [Orange] Managing Director, [Green] Director, [Blue] in House Lawyer, [Navy] Manager, [Purple] Trader, [Pink] Employee, [Gray] N/A, [Aquamarine] Not Enron's Employee.

TABLE III. FINDING IMPORTANT NODES USING FORMULA 2

Rank	Name	Designation at Enron
1	Jeffrey Skilling	CEO
2	Kevin Hyatt	Director
3	Phillip Platter	Employee
4	Mark Guzman	Managing Director
5	Matt Motley	Director

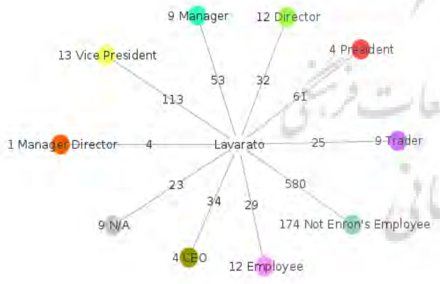

Fig. 3. Lavarato's first-level neighbor. Numbers on edges indicate number of emails sent by Lavarato to the people distinct by their Designations.

In *Table 4*, surprisingly, a simple employee called Jeff Dasovich discovered among other important persons! By investigating this person in details using additional information, it was found that Mr. Dasovich had a very important role in communication between Enron Company and government.

Discovering important persons that have ordinary positions in the company could be considered as one of the advantage of this method. The complexity of our proposed method is *O(m)* which *m* indicates the number of edges.

TABLE IV. FINDING IMPORTANT NODES USING OUR PROPOSED METHOD

Rank	Name	Designation at Enron
1	John Lavorato	CEO
2	Jeff Dasovich	Employee
3	Mike Grigsby	Manager
4	Lynn Blair	Director
5	Kenneth Lay	CEO

Clearly, in spars graphs complexity becomes *O(n)* which *n* is the number of nodes and in dense graphs complexity becomes near to $O(n^2)$.

3-3. Comparing With Existing Methods

In this section, the results of this paper were compared with 2 categories of methods. The first category contains methods that use general measures defined in graph theory for discovering the important nodes. And the second category uses specific algorithms on the Enron datasets. As an example of method for the first category, degree and betweenness centrality were used as measurements.

Graph is considered as undirected and its nodes were sorted out descendingly, according to their degree. *Table 5* shows 5 persons with the highest degree.

Another method in the first category is Betweenness centrality [46] which is calculated according to (7).

$$C_v = \sum_{s,t \neq v} \frac{\Omega_v(s,t)}{\Omega(s,t)} \qquad (7)$$

In (7), Ω (s, t) is the number of shortest paths between s and t and Ωv (s, t) is the number of shortest paths between s and t that goes through v. Table 6 shows 5 persons with

TABLE V. FINDING IMPORTANT NODES USING NODE'S DEGREE

Rank	Name	Designation at Enron	Degree
1	Jeff Dasovich	Employee	76,812
2	Tana Jones	N/A	28,945
3	James Steffes	Vice President	22,358
4	Richard Shapiro	Vice President	20,406
5	Sara Shackleton	N/A	19,737

highest score according to betweenness centrality measure. *Fig. 4* shows Bill Williams and his first-level neighbors.

Our results these methods with respect to both discovering more important nodes: CEO in our method comparing to broker in the centrality Betweenness method and Complexity: *O(n)* vs $O(n^2)$.

From the second category, we compare our result with Shetty at al [42] which report its result on the same dataset. *Table 7* shows the 5 important persons discovered by Shetty et al [42].

Comparing to Shetty et al [42], first, our proposed method discover john Lavorato with CEO position as most important person but Shetty et al [42] discover Louise Kitchen with president position as important node. Second, our proposed method succeed to discover 2 out of 4 CEOs but Shetty et al [42] discover only one CEO. Third, There is a person with ordinary employee position among 5 discovered important persons by Shetty et al [42] who does not have any skill or an important role (according to our investigation) but ordinary employee discovered by our proposed method has important role in communicating with the governments. *Table 8* shows only the most important person that discovered by different methods.

4. CONCLUSIONS AND FUTURE WORK

Recently discovering important nodes in graph data attracted a lot of attention. For example, in biology, it has been shown that important proteins are more involved in cancer or disease [47,48]. In terroristic networks, we can consider leaders as important nodes [49]. In this paper, we consider importance of people as their position in company. For finding leaders in this company, a new method was introduced based on normal entropy. Furthermore structural features of graph data such as degree were used in addition to inherent features of social network such as time of conversation between two people. Comparing the results with previous works the method proposed in this paper, seems more efficient for discovering leaders of company.

Regarding future research, we see two particularly important directions for refinement and extension of our approach. Firstly, a more advanced text mining methods to analysis the emails' contents can be used. Secondly, the proposed method is applicable to other domains such as biology.

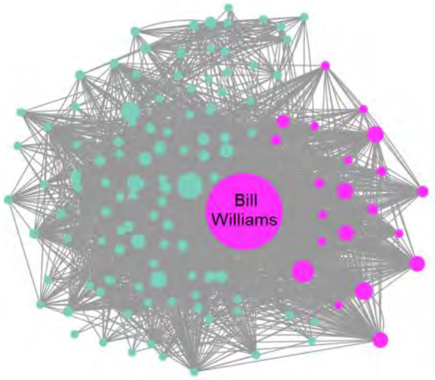

Fig. 4. Williams's first-level neighbors. Pink nodes are Enron's employees and blue nodes are others. nodes' size indicates centrality betweenness.

TABLE VI. FINDING IMPORTANT NODES USING FORMULA 7

Rank	Name	Designation at Enron
1	Bill Williams	Broker
2	Steven Merris	N/A
3	Eric Linder	Employee
4	Kay mann	Employee
5	Louise Kitchen	President

TABLE VII. FINDING IMPORTANT NODES USING SHETTY ET AL [42] PROPOSED METHOD

Rank	Name	Designation at Enron
1	Louise Kitchen	President
2	Mike Grigsby	Manager
3	Greg Whalley	President
4	Scott Neal	Employee
5	Kenneth Lay	CEO

TABLE VIII. COMPARING MOST IMPORTANT NODES FOUND BY DIFFERENT METHODS

Method	Name	Designation at Enron
Our proposed method	John Lavorato	CEO
Shetty (length 1) [42]	Louise Kitchen	President
Centrality betweenness [46]	Bill Williams	Broker
Degree	Jeff Dasovich	Employee

REFERENCES

[1] K. Yada, H. Motoda, T. Washio, and A. Miyawaki, "Consumer behavior analysis by graph mining technique," *New Mathematics and Natural Computation*, vol. 2, no. 01, pp. 59-68, 2006.

[2] S. Parthasarathy, S. Tatikonda, and D. Ucar, "A survey of graph mining techniques for biological datasets," *Managing and mining graph data*, pp. 547-580: Springer, 2010.

[3] H. Rahmani, H. Blockeel, and A. Bender, "Using a human drug network for generating novel hypotheses about drugs," *Intelligent Data Analysis*, vol. 20, no. 1, pp. 183-197, 2016.

[4] G. Klopman, "Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules," *Journal of the American Chemical Society*, vol. 106, no. 24, pp. 7315-7321, 1984.

[5] C. Phua, V. Lee, K. Smith, and R. Gayler, “A comprehensive survey of data mining-based fraud detection research,” *arXiv preprint arXiv:1009.6119*, 2010.

[6] L. Xiaojun, H. Song, and X. Weikun, “The Analysis of Logistics Influence of the Important Node Cities of Beijing-Tianjin-Hebei,” *International Journal of Business and Economics Research,* vol. 6, no. 5, pp. 88, 2017.

[7] V. D. Blondel, A. Decuyper, and G. Krings, “A survey of results on mobile phone datasets analysis,” *EPJ Data Science*, vol. 4, no. 1, pp. 10, 2015.

[8] L. Garton, C. Haythornthwaite, and B. Wellman, “Studying online social networks,” *Journal of Computer‐ Mediated Communication,* vol. 3, no. 1, pp. 0-0, 1997.

[9] R. A. Hanneman, and M. Riddle, "Introduction to social network methods," University of California Riverside, 2005.

[10] S. Wasserman, and K. Faust, *Social network analysis: Methods and applications*: Cambridge university press, 1994.

[11] W.-S. Yang, J.-B. Dia, H.-C. Cheng, and H.-T. Lin, "Mining social networks for targeted advertising." pp. 137a-137a.

[12] P. Kazienko, and K. Musial, “On utilising social networks to discover representatives of human communities,” *International Journal of Intelligent Information and Database Systems,* vol. 1, no. 3-4, pp. 293-310, 2007.

[13] A. Culotta, R. Bekkerman, and A. McCallum, “Extracting social networks and contact information from email and the web,” *Computer Science Department Faculty Publication Series*, pp. 33, 2004.

[14] L. A. Adamic, and E. Adar, “Friends and neighbors on the web,” *Social networks,* vol. 25, no. 3, pp. 211-230, 2003.

[15] P. Kazienko, and K. Musiał, "Mining personal social features in the community of email users." pp. 708-719.

[16] F. Kooti, L. M. Aiello, M. Grbovic, K. Lerman, and A. Mantrach, "Evolution of conversations in the age of email overload." pp. 603-613.

[17] I. Alsmadi, and I. Alhami, “Clustering and classification of email contents,” *Journal of King Saud University-Computer and Information Sciences,* vol. 27, no. 1, pp. 46-57, 2015.

[18] L. Getoor, E. Segal, B. Taskar, and D. Koller, "Probabilistic models of text and link structure for hypertext classification." pp. 24-29.

[19] L. C. Freeman, “Centrality in social networks conceptual clarification,” *Social networks,* vol. 1, no. 3, pp. 215-239, 1978.

[20] M. Kaur, and S. Singh, “Analyzing negative ties in social networks: A survey,” *Egyptian Informatics Journal,* vol. 17, no. 1, pp. 21-43, 2016.

[21] C. C. Noble, and D. J. Cook, "Graph-based anomaly detection." pp. 631-636.

[22] T. Kajdanowicz, and M. Morzy, “Using graph and vertex entropy to compare empirical graphs with theoretical graph models,” *Entropy,* vol. 18, no. 9, pp. 320, 2016.

[23] S. White, and P. Smyth, "Algorithms for estimating relative importance in networks." pp. 266-275.

[24] M. Newman, “Who is the best connected scientist,” *A Study of Scientific Coauthorship Networks, Santa Fe Institute, Santa Fe, NM*, pp. 1-32, 2000.

[25] K. Zhang, H. Zhang, Y. dong Wu, and F. Bao, "Evaluating the importance of nodes in complex networks based on principal component analysis and grey relational analysis." pp. 231-235.

[26] S. Huang, H. Cui, and Y. Ding, “Evaluation of node importance in complex networks,” *arXiv preprint arXiv:1402.5743*, 2014.

[27] C. Gao, D. Wei, Y. Hu, S. Mahadevan, and Y. Deng, “A modified evidential methodology of identifying influential nodes in weighted networks,” *Physica A: Statistical Mechanics and its Applications,* vol. 392, no. 21, pp. 5490-5500, 2013.

[28] J. Wang, L. Rong, and T. Guo, “A new measure method of network node importance based on local characteristics,” *Journal of Dalian University of Technology,* vol. 50, no. 5, pp. 822-826, 2010.

[29] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, “Identifying influential nodes in complex networks,” *Physica a: Statistical mechanics and its applications,* vol. 391, no. 4, pp. 1777-1787, 2012.

[30] K. Saito, M. Kimura, K. Ohara, and H. Motoda, “Super mediator–A new centrality measure of node importance for information diffusion over social network,” *Information Sciences,* vol. 329, pp. 985-1000, 2016.

[31] S. Xue, L. Xiong, Z. Lu, and J. Wu, “Graph-theoretic node importance mining in world city networks: methods and applications,” *Information Discovery and Delivery,* vol. 45, no. 2, pp. 57-65, 2017.

[32] B. Klimt, and Y. Yang, "The enron corpus: A new dataset for email classification research." pp. 217-226.

[33] M. Dehmer, and A. Mowshowitz, “A history of graph entropy measures,” *Information Sciences,* vol. 181, no. 1, pp. 57-78, 2011.

[34] J. Körner, “Fredman–Komlós bounds and information theory,” *SIAM Journal on Algebraic Discrete Methods,* vol. 7, no. 4, pp. 560-570, 1986.

[35] J. Kieffer, and E.-h. Yang, “Ergodic behavior of graph entropy,” *Electronic Research Announcements of the American Mathematical Society,* vol. 3, no. 2, pp. 11-16, 1997.

[36] L. Ming, and P. Vitányi, *An introduction to Kolmogorov complexity and its applications*: Springer Heidelberg, 1997.

[37] H. Buhrman, M. Li, and P. Vitanyi, "Kolmogorov random graphs." pp. 78-96.

[38] Xycoon. "Statistics - Econometrics - Forecasting," 2018-03-13; https://www.xycoon.com/normalized_entropy.htm.

[39] E. Data. "EnronData," 2018-03-13; http://www.enrondata.org.

[40] R. Rowe, G. Creamer, S. Hershkop, and S. J. Stolfo, "Automated social hierarchy detection through email network analysis." pp. 109-117.

[41] P. Kazienko, K. Musial, and A. Zgrzywa, “Evaluation of node position based on email communication,” *Control & Cybernetics,* vol. 38, no. 1, 2009.

[42] J. Shetty, and J. Adibi, "Discovering important nodes through graph entropy the case of enron email database." pp. 74-81.

[43] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park, “Scan statistics on enron graphs,” *Computational & Mathematical Organization Theory,* vol. 11, no. 3, pp. 229-247, 2005.

[44] A. H. Ruhe. "Enron Mail Database," 2018-03-13; http://www.ahschulz.de/enron-email-data/.

[45] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, “Cytoscape: a software environment for integrated models of biomolecular interaction networks,” *Genome research,* vol. 13, no. 11, pp. 2498-2504, 2003.

[46] M. E. Newman, "Who is the best connected scientist? A study of scientific coauthorship networks," *Complex networks*, pp. 337-370: Springer, 2004.

[47] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, and S. Koeppen, “A human protein-protein interaction network: a resource for annotating the proteome,” *Cell,* vol. 122, no. 6, pp. 957-968, 2005.

[48] H. Rahmani, H. Blockeel, and A. Bender, "Predicting the functions of proteins in protein-protein interaction networks from global information." pp. 82-97.

[49] V. Krebs, “Connecting the dots–social network analysis of 9-11 terror network,” *Last modified*, 2008.

Vahid Bashiri is fourth year bachelor students in computer engineering department of Iran Univerity of Science and Technology in 2018. His main research areas are data mining and machine learning.

Hossein Rahmani received his Ph.D. in computer science from Leiden university, the Netherlands, in 2012. From 2012 to 2015, he worked as a postdoctoral researcher at the University of Maastricht, the Netherlands. In September 2015, he joined the computer engineering department of Iran university of science and technology. His research interests are data/text/graph mining, machine learning and network analysis.

Hamid Bashiri is fourth year bachelor students in computer engineering department of Iran Univerity of Science and Technology in 2018. His main research areas are data mining and machine learning.