DOI: 10.37943/AITU.2021.99.34.007

DOI: 10.37943/AITU.2021.99.34.007

A. Kalikova
MA, Senior-lecturer
aigerim.kalikova@astanait.edu.kz, orcid.org/0000-0003-4429-0006
Astana IT University, Kazakhstan
Monash University, Australia

STATISTICAL ANALYSIS OF RANDOM WALKS ON NETWORK

Abstract: This paper describes an investigation of analytical formulas for parameters in random walks. Random walks are used to model situations in which an object moves in a sequence of steps in randomly chosen directions. Given a graph and a starting point, we select a neighbor of it at random, and move to this neighbor; then we select a neighbor of this point at random, and move to it etc. It is a fundamental dynamic process that arises in many models in mathematics, physics, informatics and can be used to model random processes inherent to many important applications. Different aspects of the theory of random walks on graphs are surveyed. In particular, estimates on the important parameters of hitting time, commute time, cover time are discussed in various works. In some papers, authors have derived an analytical expression for the distribution of the cover time for a random walk over an arbitrary graph that was tested for small values of n. However, this work will show the simplified analytical expressions for distribution of hitting time, commute time, cover time for bigger values of n. Moreover, this work will present the probability mass function and the cumulative distribution function for hitting time, commute time.
Keywords: random work, hitting time, commute time, cover time

Introduction

Random walks arise in many models in mathematics and physics. In fact, this is one of those notions that tend to pop up everywhere once we begin to look for them. It can explain the observed behaviors of many processes in scientific fields, such as computer science, physics, ecology, biology, economics and chemistry.

The aim of this paper is to show a result of an investigation of simple, exact formulas for three important quantities in random walks: hitting time, commute time and cover time for big values of n and their and simulation result.

A random walk on a graph is a very simple discrete time process. A particle starts moving on the vertices of the graph [1,2,3]. It starts at a specific vertex and at each time step it moves from its present vertex to one of its neighbours that are chosen at random. The measures in the quantitative theory of random walks that are main parameters of this paper are: the *hitting time* is a number of steps before node *j* is first time visited, starting from node *i*; the *commute time:* this is a number of steps in a random walk starting at *i*, before node *j* is visited and then node *i* is reached again; the *cover time* (starting from a given distribution) is a number of steps to reach every node [1,2].

Various aspects of the theory of random walks on graphs are surveyed. There can be found different works devoted to random walks. From those papers, more relevant will be analyzing [4,5,6] random walks on complex networks with derived an expression for hitting time

Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

between two nodes and [7,8] the properties of random walks on complex trees that both will help to understand concept and develop them. Another theoretical literature that has been reviewed is [1] that gives main direction for developing analytical expressions of random walk's parameters with some given examples of special cases for different graphs.

Outline

In this paper, an investigation of deriving formulas and statistical simulation of three main parameters: hitting time, commute time and cover time will be shown. Firstly, there will be presented introduction of main properties and parameters of random walk through connections with the eigenvalues of graphs by analyzing existing, relevant work. Secondly, simulations of random walk on MATLAB Simulink and there will be shown statistical analysis.

Theory

Given a graph and a starting point, we select a neighbour of it at random, and move to this neighbour; then we select a neighbour of this point at random, and move to it etc. The (random) sequence of points selected this way is a *random walk* on the graph.

Let $G = (V,E)$ be a connected graph with n nodes and m edges. We consider a random walk on G: by starting at an initial vertex v_0; we move to its neighbours with probability $1/d(v_t)$, where d is a degree of the vertex. After number of step t we are at a node v_t, Clearly, the sequence of random nodes $(v_t : t = 0; 1; 2; 3 \ldots n:)$ is a Markov chain. The node v_0 may be fixed, but may itself be drawn from some initial distribution P_0 [4,9].

We denote by P_t the distribution of v_t:

$$P_t = Prob\ (v_t = i) \tag{1}$$

It is a stochastic matrix with p_{ij} for all $i,j \in G$ and $\sum_{j \in G} p_{ij} = 1$

So

$$p_{ij} = \begin{cases} \frac{1}{d(i)}, if\ ij \in G, \\ 0, otherwise. \end{cases} \tag{2}$$

Now we can consider the n-step transition matrix. Let denote the event E that has a probability $P_i(X_1=j)=p_{ij}$ and using initial distribution π we get:

$$P_i(X_1 = j) = \sum_{i \in G} \pi_i P_i(X_1 = j) = \sum_{i \in G} \pi_i p_{ij},$$

similarly

$$P_i(X_2 = j) = \sum_{i,k} \pi_i P_i(X_1 = k, X_2 = j) =$$

$$= \sum_{i,k} \pi_i p_{ik} p_{kj} = P_{ij}^2, \tag{3}$$

by calculating in this way, we get

$$P_i(X_n = j) = (P^n)_{ij} = p_{ij}^{(n)} \tag{4}$$

$$P(X_n = j) = (\pi P^n)_j \tag{5}$$

Thus, we obtain the simple equation in n-step transition matrix that we will use in our further derivation:

DOI: 10.37943/AITU.2021.99.34.007

$$P^{(n)} = p^{(n)}{}_{ij}, \tag{6}$$

where P^n is raised to power n

Moreover, for all i, j and $n, m \geq 0$, hold Chapman-Kolmogorov equations:

$$p_{ij}^{(n+m)} = \sum_{k \in G} p_{ik}^{(n)} \, p_{kj}^{(m)} \tag{7}$$

We will review the most relevant classification of States of a Markov Chain for this work:

• Absorbing: A state is said to be an absorbing state if, upon reaching this state, the process never will leave this state again. State i is an absorbing state if and only if $p_{jj} = 1$.

• Accessible: A state j is said to be accessible from state i if $p(n)_{ij} > 0$ for some $n \geq 0$ i.e. one can get to j from i in some steps.

• Communicating: If state j is accessible from state i and state i is accessible from state j then states i and j are said to communicate [5,10].

Part 1. Hitting time

The hitting time is a number of steps before node j is first time visited, starting from node i. We denote it as H.

In order to calculate hitting time, we start to get a new matrix A_j, which is our Initial transition matrix P, but is replaced with one absorbing state. A_j will have $a_{jj}=1$ and $a_{ij}=0$ for all $i \neq j$. By using it we can calculate steps until reaching that absorbing state, where process will never leave that state again [1].

We can denote $A_j = (a_{ij})_{i,j \in G}$, where a_{ij} is ij the entry of the matrix A_j. The probability to get from starting node i to absorbing state j in t step (using power t) is $a^t{}_{ij}$ that give us already cumulative distribution of hitting time (H) denoting as F.

To calculate probability distribution of mass function of hitting time (H) we use subtraction of CDF of H on one power (step) difference:

$$p_H(t) = a_{ij}^{(t)} - a_{ij}^{(t-1)} \tag{8}$$

Example:

There can be illustrated a simple example of calculation of hitting time by using Markov chain transition matrix. Lets take an arbitrary graph with 3 nodes:

$$P_{ij} = \begin{vmatrix} 0.2 & 0.3 & 0.2 \\ 0.3 & 0 & 0.1 \\ 0.2 & 0.1 & 0 \end{vmatrix}$$

So, we get 3×3 matrix and to calculate hitting time, we need to replace one node with absorbing one, let make state 2 is absorbing. Then we get a new A matrix that is a modified matrix P, but with an absorbing node 2.

$$A_{ij} = \begin{vmatrix} 0.2 & 0.3 & 0.2 \\ 0.3 & 0 & 0.1 \\ 0.2 & 0.1 & 0 \end{vmatrix}$$

Let's now see the probability transition matrix in 30 step, by starting with node 1 to be absorbed in state 2, we get cumulative distribution function at 30 step, $a_{12}^{(30)}$, if we get one more step $a_{ij}^{(t+1)} = a_{12}^{(31)}$.

Now we can calculate probability mass function of hitting time of matrix A as following:

$$p_H(t) = a_{ij}^{(t)} - a_{ij}^{(t-1)} = a_{12}^{(31)} - a_{12}^{(t-30)}$$

Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

By running simulation for 30 nodes in MATLAB, we see the probability mass function (PMF) and cumulative distribution function (CDF) of hitting time. From the Figure 1, it can be seen that the result of simulation of PMF is quite similar to the equation result, while simulation and equation in CDF lines are completely similar Figure 2.

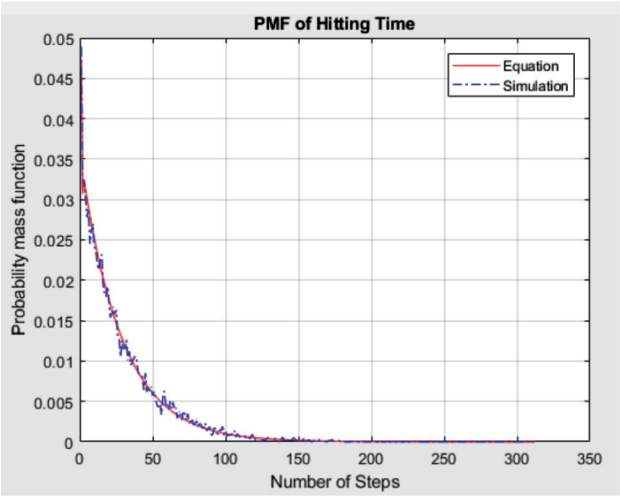

Fig. 1. PMF of hitting time for 30 nodes

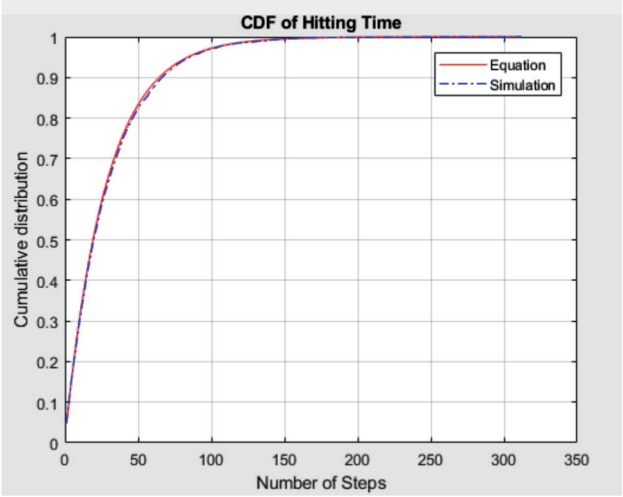

Fig. 2. CDF of hitting time for 30 nodes

Now, by running simulation for 100 nodes in MATLAB, we can observe that the probability mass function (PMF) and cumulative distribution function (CDF) of hitting time have changed. From the Figure 3, it can be seen that the result of simulation of PMF has some disturbance compared to the equation result. However, CDF result of simulation and equation has stayed the same (Figure 4).

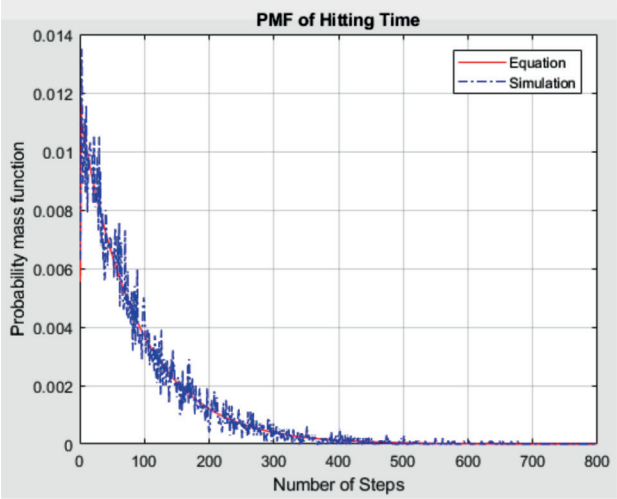

Fig. 3. PMF of hitting time for 100 nodes

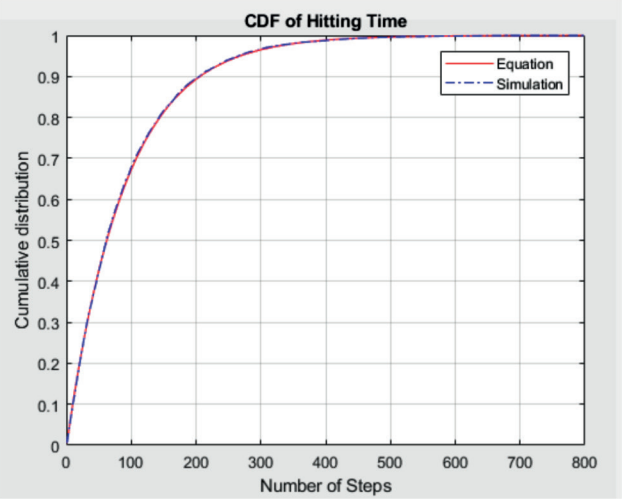

Fig. 4. CDF of hitting time for 100 nodes

Let's increase number of nodes to 300, we can see that probability mass function (PMF) of hitting time has changed. From the Figure 5, it can be seen that the result of simulation of PMF has more disturbance compared to the equation result, but it still has a similar trend with equation result. In these plots, number of steps are approximately 3500. From Figure 6, it can be seen a cumulative distribution function (CDF) of hitting time for this number of nodes.

DOI: 10.37943/AITU.2021.99.34.007

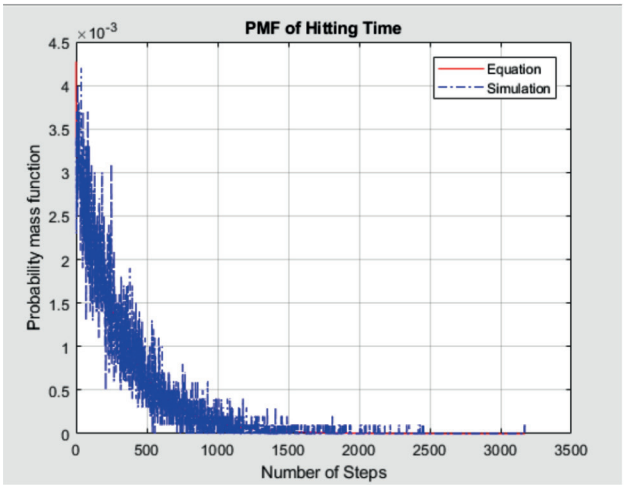

Fig. 5. PMF of hitting time for 300 nodes

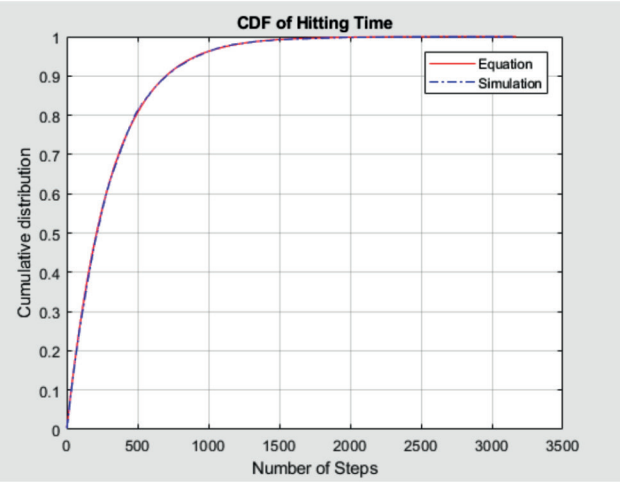

Fig. 6. CDF of hitting time for 300 nodes

Further in calculating cover time, we need also to use unions of hitting time. Let denote event as E and reaching node v from staring node that $i \neq v$. Then event by time t is E^t_{vk} and sequence of events are E^t_{v1}, E^t_{v2}, $E^t_{v3,\dots,}E^t_{vn}$. To calculate hitting time of union of those events, we replace nodes v_1, v_2,v_3, …,v_n with absorbing states as we did before for single row, but this case for each event. We get then cumulative distribution by summing the events union hitting time:

$$F_{union\ H}(t) = \sum_{v=1}^{n} a^t_{ik} \tag{9}$$

Now, we want to get probability mass function of hitting time (H) of the union events. Similarly as we did with single hitting time, but here we use CDF of the union of hitting time, we get the fallowing:

$$p_{union\ H} = F_{union\ H}(t) - F_{union\ H}(t-1), \tag{10}$$

this equation we use further in calculation of cumulative distribution function of the cover time [1, 11].

Part 2. Commute time

The commute time is a number of steps in a random walk starting at i, before node j is visited and then node i is reached again. We had hitting time as $H(i,j) \neq H(j,i)$, but for commute time we have a symmetrical parameter, $k\ (i,l)$:

$$\mathrm{k} = H(i,j) + H(j,i) \tag{11}$$

We use here mathematical operation of two function to obtain third one, which is a modified version of one the original function – we use convolution. We give the integral of the multiplication of two functions as a function of the number of the original one:

$$p_k(t) = \sum_{\tau=1}^{t} p_{H(i,j)}(\tau) p_{h(j,i)}(t-\tau) \tag{12}$$

In order to compute cumulative distribution function of the commute time we copy our Markov chain and we get initial Markov chain and delete all outgoing edges of the node j, we modify the original. Then we change the copied Markov chain by replacing all outgoing edges of the node i' that is a copy of the node i of the initial Markov chain with a self-loop.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

Then we do a connection between the two chains by adding one directed edge from node j to its copy j' of the copied chain. Let O be $n \times n$ matrix of all zeros, O_j is to be the n×n matrix for which all elements are zeros and $O_{jj} = 1$. Next, we get initial matrix P and modify it jth row to be zeros and denote is as Ajo and we obtain a new matrix:

$$\begin{bmatrix} A_{j_0} & O_j \\ O & A_i \end{bmatrix} \tag{13}$$

We denote this matrix as M and we get cumulative distribution for commute time k-1 is the element of matrix $m_{i,i+n}^t$ which is the element of M^t.

By running simulation for 10 nodes in MATLAB, we see probability mass function (PMF) and cumulative distribution function (CDF) of commute time. From the Figure 7, it can be seen that the result of simulation of PMF is approximately 0.11 and number of steps more than 135. From figure 8, it can be seen CDF of commute time for 10 nodes and it starts from 0.1.

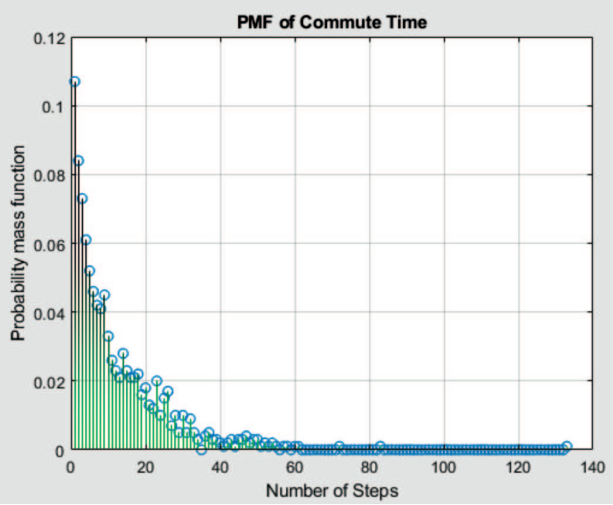

Fig. 7. PMF of commute time for 10 nodes

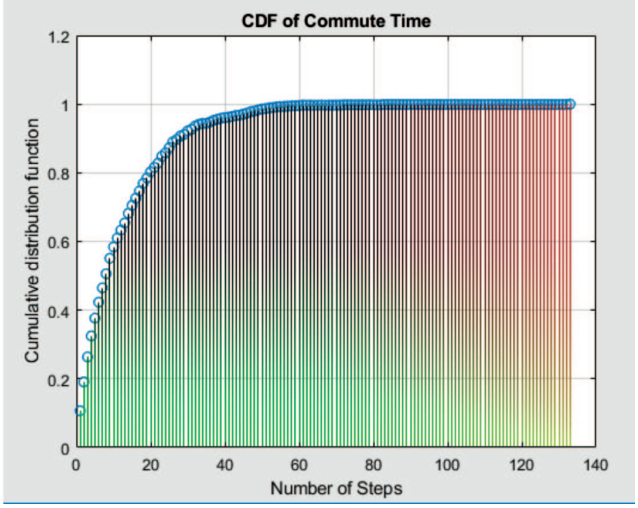

Fig. 8. CDF of commute time for 10 nodes

Part 3. Cover time

The cover time is a number of steps, starting from a given distribution, to reach every node in the graph.

We start with considering the inclusive and exclusive principle of probabilities of events [1, 12]:

$$A \cup B = A + B - A \cap B, \tag{14}$$

where a general form is

$$\bigcup_{i=1}^{n} A_i = \sum_{i=1}^{n} A_i - \sum_{1 \le i < j < k \le n} A_i \cap A_j + \cdots + \sum_{1 \le i < j < k \le n} A_i \cap A_j \cap A_k - \cdots + (-1)^{n-1}(A_1 \cap \ldots \cap A_n)$$

In our case, we use union of events to calculate intersection of nodes. We can combine the principle of inclusion and exclusion with Morgan's theorem that can be used to count intersection of sets as well, we have:

$$\bigcap_{i=1}^{n} A_i = \overline{\bigcup_{i=1}^{n} \overline{A_i}} \tag{15}$$

With that we can turn the problem to find intersection or union.

We already know how to get a union of the events – hitting time of the events $p_{(union\ H)}$ that we derived a formula before. And to calculate cumulative distribution of the cover time we need also CDF of hitting time to each node in the graph, which also we have derived.

DOI: 10.37943/AITU.2021.99.34.007

Now we can get the formulas of F_c cumulative distribution function of the cover time by using hitting time and union of events of hitting time, where s is a starting node:

$$F_c(t) = \sum_{i=1, i\neq s}^{n} F_{v_i} - \sum_{i=1, i\neq s}^{n} \sum_{j=i+1, j\neq s}^{n} F_{v_i v_j} + \ldots + (-1)^{n-1} F_{v_1, v_2, \ldots, v_n}(t) \quad (16)$$

Then we can find probability mass function from the CDF of cover time by:

$$p_c = F_c(t) - F_c(t-1) \quad (17)$$

Conclusion

Random Walks are used to model situations in which an object moves in a sequence of steps in randomly chosen directions. Random walks are one of the basic objects studied in probability theory. The motivation comes from observations of various random motions in physical and biological sciences. Many phenomena can be modeled as a random walk and we have seen several examples in this work. This work has shown an investigation and derivation of simple, exact formulas for three important quantities in random walks: hitting time, commute time and cover time for big values of n. There were presented and explained theory, equation on a random walk and results of the simulations of random walk on MATLAB Simulink.

References

1. Zlatanov, N., & Kocarev, L. (2009). Random walks on networks: Cumulative distribution of cover time. *Physical Review E,* 80(4), 041102
2. Kang, U., Tong, H., & Sun, J. (2012, April). Fast random walk graph kernel. *In Proceedings of the 2012 SIAM international conference on data mining* (pp. 828-838). Society for Industrial and Applied Mathematics.
3. Adeleye, O., Yu, J., Ruan, J., & Sheng, Q. Z. (2020, February). Evaluating random walk-based network embeddings for web service applications. *In Australasian Database Conference* (pp. 198-205). Springer, Cham.
4. Amancio, D.R., Silva, F.N., & Costa, L. D.F. (2015). Concentric network symmetry grasps authors' styles in word adjacency networks. *EPL (Europhysics Letters),* 110(6), 68001.
5. Li, X., Han, Z., Wang, L., & Lu, H. (2015). Visual tracking via random walks on graph model. *IEEE transactions on Cybernetics,* 46(9), 2144-2155.
6. Dong, X., Shen, J., Shao, L., & Van Gool, L. (2015). Sub-Markov random walk for image segmentation. IEEE Transactions on Image Processing, 25(2), 516-527.
7. Spitzer, F. (2013). Principles of random walk (Vol. 34). Springer Science & Business Media.
8. Backstrom, L., & Leskovec, J. (2011, February). Supervised random walks: predicting and recommending links in social networks. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 635-644).
9. Condamin, S., Bénichou, O., Tejedor, V., Voituriez, R., & Klafter, J. (2007). First-passage times in complex scale-invariant media. *Nature,* 450(7166), 77-80.
10. Nghiem, T.P., Maulana, K., Waluyo, A.B., Green, D., & Taniar, D. (2013, March). Bichromatic reverse nearest neighbors in mobile peer-to-peer networks. *In 2013 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (pp. 160-165). IEEE.
11. Arunachalam, A., & Sornil, O. (2015, February). An analysis of the overhead and energy consumption in flooding, random walk and gossip based resource discovery protocols in MP2P networks. *In 2015 Fifth International Conference on Advanced Computing & Communication Technologies* (pp. 292-297). IEEE.
12. Gurkan, T., Resul, D., Ramakrishnan B., Kilicaslan Y. (2017) Big Data Analysis for M2M Networks: Research Challenges and Open Research Issues. *International Journal of Computer Networks and Applications (IJCNA),* 27-34.