A Direct Proof of Significant Directed Random Walk

IOP Conf. Series: Materials Science and Engineering **235** (2017) 012004 doi:10.1088/1757-899X/235/1/012004

A Direct Proof of Significant Directed Random Walk

Choon Sen Seah[1, a], Shahreen Kasim[1, b], Mohd Farhan Md Fudzee[1, c], and Mohd Saberi Mohamad[2, d]

[1]Soft Computing and Data Mining Centre, Faculty of Computer Sciences and Information Technology, Universiti Tun Hussein Onn Malaysia
[2]Faculty of Computing, Universiti Teknologi Malaysia, Skudai

[a]hi150021@siswa.uthm.edu.my, [b]shahreen@uthm.edu.my, [c]farhan@uthm.edu.my, [d]saberi@utm.my

Abstract. This paper is presented to disclose the relationship between weight and connectivity of nodes. An equation is formed to enhance the connectivity of nodes in directed graph via weigh. With implementation of references data, the adjacency matrix is further enhances to increases the accessibility of nodes via vector. The evolution of random walk is disclosed in this paper as well. Significant directed random walk will be used to prove the importance of weight in this paper.

1. Introduction

The basics of random walk can be traced back from a famous study by Brown (1828), known as Brownian motion [1]. Random walk was first developed to predict the mosquito infestation in forest by Karl Pearson [2]. Nowadays, random walk is applied in many different fields for prediction purposes and hence, it was further developed to improve the accuracy of prediction. Random walk was developed into directed random walk where it is classifying under biased random walk [3]. Directed graph consists of a sequence of vertices with connection of edges [4]. With a specific vector, an initial vertex will be transported to another vertex via edges. Implementation of directed graph will be further discussed on section 3 and 4. The connection between vertices are improved with the implementation of weight during the calculation. The equations are further proved through implementing numeric values of gene expression data after data pre-processing via Gene Chip Robust Multiarray Averaging (GCRMA). The comparison of results between directed random walk and significant directed random walk are shown in section 5.

2. Random Walk

Random walk can be described as the movement of a particle in a certain state space under the random action [5]. The state space is usually a dimensional Euclidean space or the integral lattice. Furthermore, random walk was then drawn to the subject and many important fields, such as random processes, random noise, spectral analysis and stochastic equations [3]. Each step of random walk is either equal to +1 (step forward) or -1 (step backward).

3. Directed Random Walk

The Directed Random Walk (DRW) developed by Liu have the ability to restart and stimulate a random walker that starts on a source node, *s* [6]. At every time step, the walker transit from its current

node to another randomly selected neighbour (forward) or goes backward to source node *s* with probability *r*. Formally, the DRW with restart is defined as

$$W_{t+1} = (1-r)M^T W_t + rW_0 \quad (1)$$

where W_t is a vector which the i node holds the probability of being at node *i* at time, *t*. *M* is the row-normalized adjacency matrix of the graph, G. When the random walk begins, the initial probability vector, W_0 was constructed by assigning to each node whose initial probability was 0. W_0 is absolute t-test score, which will be further normalized into a unit vector [6]. The restart probability *r* was set as 0.7. W_t converges to a unique steady state in the presence of the ground node. This was obtained by performing the iteration until the normalization fall between W_t and $W_{t+1} < 10^{-10}$.

4. Significant Directed Random Walk

In this section, an improved DRW which named as significant directed random walk (sDRM) is presented. The sDRW proved that the weight of the nodes can affect the connectivity of nodes, which leads to higher vector. Directed graph is defined as weighted graph when there are values attached to the directed edges [7]. These values represent the cost of travelling from one node to the other. The cost can be measure in many terms, depending on the application.

For example, distance between two nodes and the average travelling time in minutes. In the case of gene classification, weight of gene is used as a parameter to identify the usefulness in classification [8]. The significant genes normally have higher weight compared to the others. This is because of the common genes across protein-protein interaction network. Weight of genes can be obtain from gene expression data.

The Significant Directed Random Walk have ability to enhance the connectivity between nodes by weight of nodes. At every time step, the walker transitions from its current node to a randomly selected neighbour (based on edge weights) or goes back to source node, s with probability *r*. *r* can vary according to the datasets due to the attraction of nodes [9]. For example, *r* can be 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, or 0.9. Formally, the sDRW with better connectivity is defined as

$$W_{t+1} = (1-r)(M)\left(\frac{N_1 + N_2}{2}\right) + rW_t \quad (2)$$

where, W_t is a vector of *i* node which is transmitted from *i*-1 node while *M* is an adjacency matrix developed from the original directed graph (with edges) to more strongly connected directed graph. As we stated in the previous section, weight plays an important role in nodes connectivity. Hence, weight of two connected nodes, N_1 and N_2 is implemented into the equation.

If the nodes have a strong connectivity towards previous nodes, then the vector from previous node towards it will be higher.

5. Direct Proof with Gene Expression Data

In this section, we present a direct proof of Significant Directed Random Walk that have better connectivity towards vertex compare to Directed Random Walk. By applying pathway data as references data and gene expression data as input data, we can obtain result of the equation. Hence, the ability of significant random walk can be proven.

Weight of the nodes are obtained from gene dataset, GSE19188 (Non-Small Cell Lung Cancer) [10], which will be processed in data pre-processing stage before implement into equation. Table 1 shows the weight of nodes after data pre-processing using Gene Chip Robust Multiarray Averaging (GCRMA).

Assume G; *V = {1,2,3,4,5}* where G represent directed graph, while *V* represent vertex. Figure 1 shows part of the biological pathway, leukocyte transendothelial migration where highlighted nodes will be used in calculation and will be simplified into Figure 2.

Table 1. Weight of each node that implement in graph, G

Nodes	1	2	3	4	5	6
Weight	2.338914	8.47301	6.1441	3.102989	11.38365	5.149393

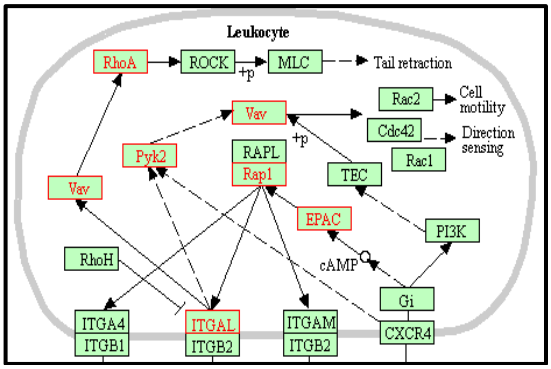

Figure 1. Gene sets that will be focusing on (highlighted)

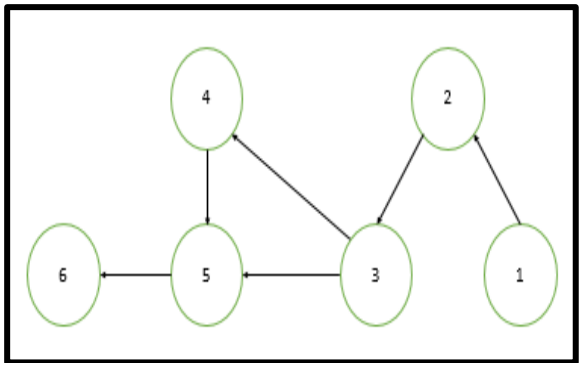

Figure 2. Graph G with nodes and edges after simplified from Figure 1

Figure 2 shows the nodes and edges that will be used in the calculation. The direction of the pathway is stated as below:
1 -> 2 -> 3 -> 4 ->5 > 6

Firstly, we will calculate the vector of significant directed random walk, followed by directed random walk, and compare the results in the end. The comparison of the results will be shown in Table 2. Initial vector, W_0 of first nodes (1) is zero because it is an initial node. Directed random walk will be proven by implementing the same data for comparison purposes. For first vector, W_t is set as 1.

Significant Directed Random Walk

W_0 = 0
W_1 = (1-0.4)(1)((2.338914+8.47301)/2) + 0.4(0)
= 3.243577
W_2 = (1-0.4)(1)((8.47301+6.1441)/2) + 0.4(3.243577)
= 4.385133 + 1.297431
= 5.682564
W_3 = (1-0.4)(1)((6.1441+3.102989)/2) + 0.4(5.682564)
= 2.774127 + 2.273026
= 5.047153
W_4 = (1-0.4)(1)((3.102989+11.38365)/2) + 0.4(5.047153)
= 4.345992 + 2.018861
= 6.364853
W_5 = (1-0.4)(1)((11.38365+5.149393)/2) + 0.4(6.364853)
= 4.959913 + 2.545941
= 7.505854

Directed Random Walk

W_0 = 0
W_1 = (1-0.4)(1)(1) + 0.4 (0)
= 0.6
W_2 = (1-0.4)(1)(0.6)+ 0.4(0)
= 0.36
W_3 = (1-0.4)(1)(0.36) + 0.4(0)
= 0.216
W_4 = (1-0.4)(1)(0.216) + 0.4(0)
= 0.1296
W_5 = (1-0.4)(1)(0.1296) + 0.4(0)
= 0.07776

Table 2. Comparison result of vector from node 1 to node 6

Vector, W	Significant Directed Random Walk	Directed Random Walk
W_0	0	0
W_1	3.243577	0.6
W_2	5.682564	0.36
W_3	5.047153	0.216
W_4	6.364853	0.1296
W_5	7.505854	0.07776

From the results above, we know the fluctuation of the vector in significant directed random walk is because of the weight. Weight plays an important role to attract the other nodes while the constant decline of vector in directed random walk is because of the previous vector. The connectivity (vector) will become weaker and weaker.

On the other hand, the connectivity between 2 nodes can also be proven by significant directed random walk. Hence, another calculation by using DRW and sDRW will be run to test the connectivity between first node and the other node. The paths are stated as below:

1 -> 2, 1 -> 3, 1 -> 4, 1 -> 5, 1 -> 6

The calculations by using sDRW and DRW are shown as below:

Significant Directed Random Walk

$W_{1->2} = (1-0.4)(1)(\frac{2.338914+8.47301}{2}) + 0.4(0)$
$= 3.243577$

$W_{1->3} = (1-0.4)(1)(\frac{2.338914+6.1441}{2}) + 0.4(0)$
$= 2.544904$

$W_{1->4} = (1-0.4)(1)(\frac{2.338914+3.102989}{2}) + 0.4(0)$
$= 1.632571$

$W_{1->5} = (1-0.4)(1)(\frac{2.338914+11.38365}{2}) + 0.4(0)$
$= 4.116769$

$W_{1->6} = (1-0.4)(1)(\frac{2.338914+5.149393}{2}) + 0.4(0)$
$= 2.246492$

Directed Random Walk

$W_{1->2} = (1-0.4)(1)(1) + 0.4(0)$
$= 0.6$

$W_{1->3} = (1-0.4)(1)(1) + 0.4(0)$
$= 0.6$

$W_{1->4} = (1-0.4)(1)(1) + 0.4(0)$
$= 0.6$

$W_{1->5} = (1-0.4)(1)(1) + 0.4(0)$
$= 0.6$

$W_{1->6} = (1-0.4)(1)(1) + 0.4(0)$
$= 0.6$

Table 3. Vector from node 1 to the other nodes

Vector, W	Significant Directed Random Walk	Directed Random Walk
$W_{1->2}$	3.243577	0.6
$W_{1->3}$	2.544904	0.6
$W_{1->4}$	1.632571	0.6
$W_{1->5}$	4.116769	0.6
$W_{1->6}$	2.246492	0.6

From Table 3, with sDRW, we figure out the connectivity between node 1 and node 5 are strongest among the other nodes, while the weakest connectivity is between node 1 and node 4. While with DRW, the connectivity is remaining the same because the initial vector and first vector play the roles in determine the next vector. In sDRW, the reason to have such significant different of vector is because of the weight between each node.

6. Conclusion

Weight plays an important role between vertex and edge. Vector can be enhanced by implementing weight as one of the parameter. The equation shown is significant directed random walk. With the proven results after implementation, we believe that connectivity of nodes can be enhances with significant directed random walk as well as improve the vector. By comparing two different random walk, we figure out that the connectivity between nodes can be determine via vector. By using vector, the direction according to the pathway are fixed and possible to be simplify. With this, the pathway data can be used as references data to enhances the accuracy of the cancerous classification. sDRW proved that enhanced pathway can increases the accessibility of nodes towards the other significant nodes. With the enhanced pathway data, the result of accuracy can be increases due to fully utilized pathway data as references data. Fully utilized references data can help in increases the accuracy of cancerous classification.

Acknowledgments
We would like to thank Universiti Tun Hussein Onn Malaysia for supporting this research under the FRGS research grants (Vot numbers: 1559), also, thanks to Gates IT Solution Sdn Bhd for the whole support.

References
[1] Woolard, E. W., Einstein, A., Furth, R., & Cowper, A. D. (1928). Investigations on the Theory of the Brownian Movement. The American Mathematical Monthly, 35(6), 318. DOI:10.2307/2298685
[2] Pearson, K. (1905). "The Problem of the Random Walk". Nature. 72 (1865): 294. DOI:10.1038/072294b0
[3] Edward A Codling, Michael J Plank, Simon Benhamou. 2008. Random walk models in biology. J. R. Soc. Interface. (August 2008), 813-834. DOI= 10.1098/rsif.2008.0014.
[4] Bapat, R., Kalita, D., & Pati, S. (2012). On weighted directed graphs. Linear Algebra and its Applications, 436(1), 99-111. DOI:10.1016/j.laa.2011.06.035
[5] Aldous and Fill, Reversible Markov Chains and Random Walks on Graphs. Retrieved February 15, 2017, from http://www.stat.berkeley.edu/~aldous/RWG/book.html.
[6] Liu, W., Li, C., Xu, Y., Yang, H., Yao, Q., Han, J., . . . Li, X. (2013). Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. Bioinformatics, 29(17), 2169-2177. DOI:10.1093/bioinformatics/btt37
[7] Bender, M. A., & Ron, D. (2002). Testing properties of directed graphs: acyclicity and connectivity. Random Structures and Algorithms, 20(2), 184-205. DOI:10.1002/rsa.10023.abs
[8] Cai H, Ruan P, Ng M, Akutsu T. Feature weight estimation for gene selection: a local hyperlinear learning approach. BMC Bioinformatics. 2014;15:70. doi: 10.1186/1471-2105-15-70.
[9] Seah, C. S., Kasim, S., and Mohamad, M. S.,"Specific Tuning Parameter for Directed Random Walk Algorithm Cancer Classification," International Journal on Advanced Science, Engineering and Information Technology, vol. 7, no. 1, 2017. [Online]. Available: http://dx.doi.org/10.18517/ijaseit.7.1.1588.
[10] Hou J, Aerts J, den Hamer B, van Ijcken W et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS One 2010 Apr 22;5(4):e10312. PMID: 20421987