ASM Sc. J., 12, 2019
https://doi.org/10.32802/asmscj.2019.400

Extracting Network Structure for International and Malaysia Website via Random Walk

Y. S. J. Liang[1], K. T. Chan[1, 2*], H. Zainuddin[1, 2] and N. M. Shah[1, 2]

[1]*Laboratory of Computational Sciences and Mathematical Physics, Institute for Mathematical Research (INSPEM), Serdang, Selangor, Malaysia*
[2]*Department of Physics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*

World Wide Web is an information retrieval system accessible via the Internet. Since all the web resources and documents are interlinks with hypertext links, it formed a huge and complex information network. Besides information, the web is also a primary tool for commercial, entertainment and connecting people around the world. Hence, studying its network topology will give us a better understanding of the sociology of content on the web as well as the possibility of predicting new emerging phenomena. In this paper, we construct networks by using random walk process that traverses the web at two popular websites, namely *google.com* (global) and *mudah.my* (local). We perform measurement such as degree distribution, diameter and average path length on the networks to determine various structural properties. We also analyse the network at the domain level to identify some top-level domains appearing in both networks in order to understand the connectivity of the web in different regions. Using centrality analysis, we also reveal some important and popular websites and domain from the networks.

Keywords: complex network; world wide web; random walk; network topology

I. INTRODUCTION

The World Wide Web (WWW) or the web is a massive information network that has been evolving ever since it was first introduced in 1989 by Tim Berners-Lee in European Organization for Nuclear Research (CERN). The web operates on a client/server model across the Internet. The core of the web is formed on two main software technologies, namely the Hypertext Markup Language (HTML) and the Hypertext Transfer Protocol (HTTP) (Silvestri, 2013). HTML is used for creating web pages and web application while HTTP is used for communication between client/server. Web resources such as documents, video and data are identified by uniform resource locators (URLs) and interlinks with hypertext links.

According to Hall and Tiropanis (2012), the web is evolving, basically, from the web of documents, in the beginning, it evolves to the web of people in early 2000 and now a web of data and social networks. This web is of utmost important because not only for the information but also as a primary tool for commercial, entertainment as well as connecting people from the whole world. Hence, the ability to study and extract hidden information from the network structure of the web will be of crucial importance in many different fields. For example, if we can accurately characterise the web content and structure, it might provide new insight to the interest of millions of individuals and organisation around the globe (Rusmevichientong *et al.*, 2001). Not only that, but information from the network structure also enhanced the design of ranking mechanisms as well as improving web crawler in terms of crawling strategy (Meusel *et al.*, 2015). In addition, studying its structure could also help detect rank manipulations such as spam

*Corresponding author's e-mail: chankt@upm.edu.my

networks, which publish large numbers of “fake" links in order to increase the ranking of a target page (Meusel *et al.*, 2015).

Since the web is very informative and important, many research works have been performed on the web such as studying its topology (Barabási, *et al.*, 2000; Broder *et al.*, 2000; Lehmberg *et al.*, 2014; Meusel *et al.*, 2015), its evolution (Hall and Tiropanis, 2012) and the different approaches in approximating certain aggregate queries of the web (Bar-Yossef *et al.*, 2000; Rusmevichientong *et al.*, 2001; Silvestri, 2013). Work by Broder *et al.* (2000) has shown that the network topology of the web can be studied using graph theoretic approaches. This information network or the web graph is a directed graph whose nodes describe the web pages while the hyperlinks are the edges. Their finding showed that the web graph has bow-tie structures and displayed power-law degree distribution. The power-law behaviour is also found from the work of Barabási *et al.* (2000), where they studied the network sampled from the *nd.edu domain*, containing 325729 nodes and 1469680 edges. Similar behaviour is also shown in work by Silvestri (2013) where he performed random walks with a jump on the web, collecting 35616 URLs in approximating 133500 steps and Meusel *et al.* (2015) where they considered the web in different aggregation level.

As the current web contains more than four billion pages and increasing each day *(www.worldwidewebsize.com)*, it is impossible to map the web completely. Besides, analysing the massive amount of data as in the work of Broder *et al.* (2000) and Meusel *et al.* (2015), it will be computationally intensive and time-consuming. However, it is still possible to work with some subgraphs of the web by utilising the random walk method as described by Silvestri (2013) aided by a programme written in Mathematica. The dataset for his work is collected using a random walk programme. Random walk is a stochastic process where the network created by taking repeated random steps (Newman, 2010). In this random walk sampling, the walker will randomly produce path by crawling to a random webpage. In his work, he has applied random walk with jumps when the walkers traverse the web. What happened here is that instead of returning to the initial URL, the walker will jump to an URL chosen at random from the walks’ history and perform additional steps. The process is repeated to generate the dataset for the web graph.

In this paper, we present a study of the structure of the web using random walk sampling method without jump-starting from two different popular websites, namely *google.com* (global) and *mudah.my* (local). *Mudah.my* is one of Malaysian favourite online marketplace selling and providing a wide variety of products and services. We expect the networks sampled from the two regions will present an informative connectivity pattern. We also studied the effect of a number of walkers on the network structure by fixing the number of steps and varying the number of random walkers. We work on the aggregation version of the network namely host-level network as this can significantly reduce the number of nodes and edges and at the same time shed some light on the connectivity of the websites rather than individual pages. We will further perform measurement such as degree distribution, diameter and average path length on the networks to determine various structural properties. We also analyse the network at domain level to identify some top-level domains appearing in both networks in order to understand the connectivity of the web in different regions. Finally, centrality analysis is used to determine the most important and influential websites in these networks.

II. MATERIALS AND METHOD

A. Sampling the World Wide Web

In this work, we utilise random walk sampling programme written in Mathematica (Silvestri, 2013) to sample the web. This is done on a 2.40 GHz Intel Core, i7-5500u CPU with 12 GB of RAM. For each website, we perform four sets of sampling, namely one walker with 20000 steps, five walkers with 4000 steps, 20 walkers with 1000 steps and 40 walkers with 500 steps. In each dataset, the total steps performed in 20000 steps. In each run, the walker will start at two fixed websites, namely *google.com* and *mudah.my*. The programme will first request the web page at the specified URL, and the server responds by sending the HTML documents. All links from the webpage will be extracted and then randomly choose a link. If the chosen webpage is unavailable, the programme will choose another link randomly. For webpages without valid outgoing links,

the programme will backtrack one step and choose again. All these processes will be repeated until all steps set by the user are exhausted. All data will be saved as text file and will be used for construction of network. According to Silvestri (2013), the programme sometimes might end earlier because all hyperlinks of the page have been exhausted or the walker is trapped in certain URL which does not have valid outgoing links. In the construction of the network, not all crawled data are used to build the network. The programme will aggregate the data by merging pages from the same hostname (e.g. *www.google.com, support.microsoft.com* etc.) into a single node. Hence, the total number of original nodes are greatly reduced, and the network formed is called the host level network.

B. Properties and Measurement of Network

1. Fundamental of network

The web is viewed as a directed network, and thus it can be represented as $G(N, E)$ where the web pages are the nodes, N while the hyperlinks are edges, E. Since we are dealing with a host-level network, the node is the hostname. The topology G can is represented by its adjacency matrix A whose elements $A_{ij} = 1(0)$ if node i and j are (not) connected by an edge. For the web network, the out-degree k_j^{out} and in-degree k_i^{in} are measured using equations $k_i^{in} = \sum_{j=1}^{n} A_{ij}$ and $k_j^{out} = \sum_{i=1}^{n} A_{ij}$ respectively. The number of edges m is just equal to the ingoing ends of edges at all nodes, or just the total number of outgoing ends of edges $m = \sum_{i=1}^{n} k_i^{in} = \sum_{j=1}^{n} k_j^{out} = \sum_{ij} A_{ij}$ (Newman, 2010). Usually, the in- and out-degree are similar for each network.

2. In- and out-degree distribution

One of the network properties of interest is the power-law degree distribution or sometimes called scale-free. Many real-world networks have shown to be a scale-free network where they have long tail probability distribution that follows the power-law expressed as $p(k) = Ck^{-\alpha}$ where $p(k)$ is the probability of the degree k, C is constant and α is the exponent. The web has been reported by Barabási *et al.* (2000), Broder *et al.* (2000) and Newman (2010) to be a scale-free network and has α a range $2 \leq \alpha \leq 3$ (Clauset *et al.*, 2009; Newman, 2010). This feature can be spotted easily when we look at the histogram of degree distribution on a log-log plot or construct the cumulative distribution function (CDF). CDF for in-degree is defined as $p_k^{in} = \sum_{j=k}^{k_{max}\Sigma} p_j^{in}$ where p_k^{in} is the probability of in-degree of k, k_{max} is the maximum of k and p_j^{in} is the probability of in-degree of j. Same goes to out-degree, a similar expression for out-degree CDF p_k^{out} can be shown as $p_k^{out} = \sum_{j=k}^{k_{max}\Sigma} p_j^{out}$.

A generally accurate method of estimating the exponent of the power-law in CDF has been explained in work by Clauset *et al.* (2009). The method used by them is the maximum likelihood estimators (MLE) which give accurate estimates in the limit of large sample size and given an appropriate minimum bound k_{min}. According to them, by setting $k_{min} \geq 6$ and $M \geq 50$, a reliable exponent which is accurate to about 1% can be achieved. Setting k_{min} as 6 for the MLE calculation is a good approximation for the power-law behaviour as the "tail" of the distribution start there. The MLE is given as

$$\hat{\gamma} = 1 + M\left[\sum_{i=1}^{M\Sigma} ln\frac{k_i}{k_{min}}[]^{-1}\right], \quad (1)$$

where k_{min} is the minimum degree for which the power-law holds and M is the number of nodes with the degree $\geq k_{min}$. For standard deviation on $\hat{\gamma}$, it is given as

$$\sigma = \sqrt{M}\left[\sum_{i}^{\Sigma} ln\frac{k_i}{k_{min}}[]^{-1}\frac{\hat{\gamma}-1}{\sqrt{M}}\right]. \quad (2)$$

Using Equations (1) and (2), we estimate the power-law exponents and the results are shown in Table 2. Networks G1 (in- and out-) and M2 (out-) do not lie in between the range $2 \leq \alpha \leq 3$ is likely due to there are quite a number of vertices with a higher degree, and also the tail does not go far as N is not big enough. This issue can be solved if

more nodes or steps are considered.

3. Centrality analysis on Host Level Network

The importance or the prestige of a website can be determined by using centrality analysis. We apply degree centrality (DC) and PageRank (PR) to capture these properties. In a directed network, DC can be defined as the number of edges incident or exiting upon a node. To justify an important node, it is more illuminating if edges are directed to it in comparison to edges exiting the node. In other words, if more webpages put a hyperlink to a website, this means that that website must be of some importance. Hence, we measure DC based on ingoing edges. As for PageRank, it is a popular measure for prestige or influence of a website. If any website is hyperlinked by many important websites, this means that the website must be of high importance too. In matrix term, PR is expressed as $x = \alpha A^T Dx + \beta$ where D is the diagonal matrix with element $D_{ii} = max(k_i^{out}, 1)$, α and β are positive constant and A is the adjacency matrix of the network (Newman, 2010).

III. RESULT AND DISCUSSION

A. Construction and Statistical Properties of the Network

Construction of the networks from the datasets is shown in Figures 1 and 2, while the statistical properties of the networks are shown in Table 1. From Figures 1 and 2, we find that the networks are getting more complicated as the number of walkers increased even though each run is set at 20000 steps. The number of nodes and edges for the networks increase as walkers increased. This increase is because, with more walkers, there is a higher probability that new paths or routes are explored thus covering more hosts in comparison to a single walker. This is confirmed by the total number of steps performed by the walkers as shown in Table 1. It shows an incremental trend except for network M2 where it shows a higher step performed as compared to M3. From the dataset of M2, it was revealed that two of its walkers have got trapped on URLs that have a lot of similar links causing them to move around the similar hosts. Since the walkers are unable to get out of these pages, each walker in M2 which can walk 4000 steps, can easily add up the performed walk but leaving the number of nodes and edges constantly. This situation can be changed if we perform a different run. From Table 1, we find that most of the walkers fail to finish the walk or stop prematurely because they might have reached some dynamical webpage, URL has zero valid links, or the server cannot be reached. The best performing network is M4 where the walkers covered 61.1% of the total steps followed by G4 with 49.8%.

The complexity of the network can also be determined from the recorded crawling time. Higher crawling time usually means walkers are still actively traversing the web, and such a dynamical process will increase the number of nodes and edges. Comparing the crawling time from the global and local site, M4 took the longest time 5.85 h to produce while G4 took around 3.8 h. The difference in crawling time for both sets of a dataset can be seen from Figure 3. The possible reasons for the difference might be due to server response time and internet access speed in different regions. From the result, it suggests that local internet speed might be slower than in US.

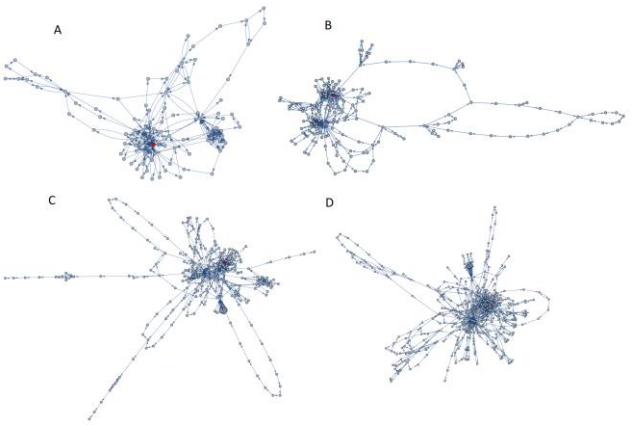

Figure 1. Networks of the web starting from *google.com* (node in red colour) with A) one walker with 20,000 steps (G1), B) five walkers with 4000 steps (G2), C) 20 walkers with 1000 steps (G3) and D) 40 walkers with 500 steps (G4)

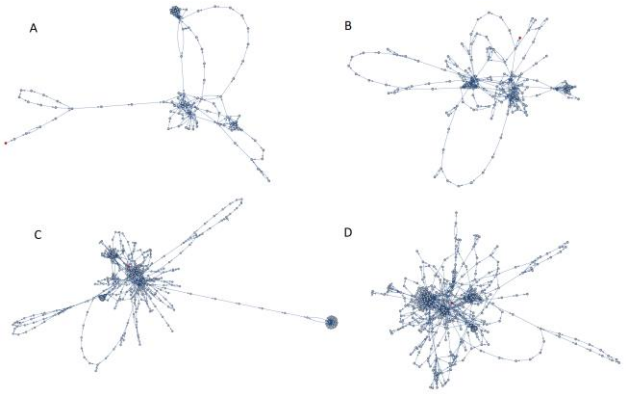

Figure 2. Networks of the web starting from *mudah.my* (node in red colour) with A) one walker with 20,000 steps (M1), B) five walkers with 4000 steps (M2), C) 20 walkers with 1000 steps (M3) and D) 40 walkers with 500 steps (M4)

Table 1. Statistical information on the networks

	Networks							
	G1	G2	G3	G4	M1	M2	M3	M4
Number of nodes	90	167	248	392	98	154	349	377
Number of edges	274	400	730	1051	309	408	899	1096
Total number of in-degree	274	400	730	1051	309	408	899	1096
Total number of out-degree	274	400	730	1051	309	408	899	1096
Mean degree	6.0889	4.7904	5.8871	5.362	6.3061	5.2987	5.1519	5.8143
Total vertex degree	548	800	1460	2102	618	816	1798	2192
Mean deviation	4.5911	3.9221	4.9973	4.646	5.4152	4.1201	4.3151	4.9101
Diameter	15	33	31	20	20	27	34	31
Average path length	4.6854	8.9444	6.5566	5.982	6.5804	7.1066	8.1726	7.0756
Density	0.0342	0.0144	0.0119	0.007	0.0325	0.0173	0.0074	0.0077
Crawling time (s)	1113.3	2147.4	10977	13669	2610.7	15014	8793.8	21077
Total walks performed	930	1178	8653	9965	896	9172	5966	12215

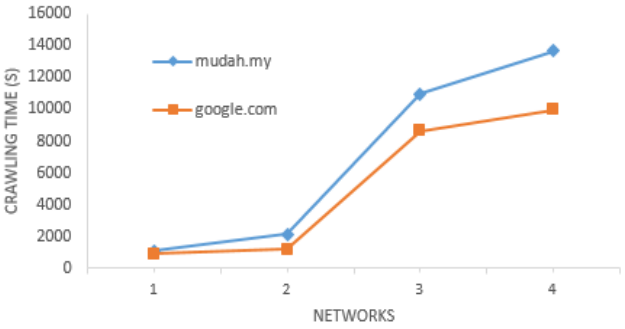

Figure 3. Crawling time for networks from *google.com* and *mudah.my*

B. Density, Diameter and Average Path Length

Network density is the ratio between actual connection to potential connection for a specified number of nodes in a network. For a directed network, this measurement is given as $m/N(N-1)$ (Newman, 2010). From Table 1, the low densities show the networks are sparse and decrease as the number of nodes increased. From the measurement, it shows lower density if the network has few edges. Sparsity of the networks here has two implications; 1) lack of RWs to sample all possible hyperlinks in each website causing many of the nodes to be periphery nodes (Csermely *et al.*, 2013) or 2) most of the websites do not have many

hyperlinks directed to them causing lower probability for the walkers to access that page again. The sparsely connected networks usually have a scale-free, power-law degree distribution (Newman, 2010).

From the connectivity of the networks, we determine the diameter and the average path length (APL) for the strongly connected component (SCC) of the networks. Strongly connected network means that for any pair of nodes u and v in the network, there is a path connecting them (Broder *et al.*, 2000). Measurement is based on the giant strong component only because if no path exists between two nodes, then the distance will be considered infinity. Noted that the giant strong component is the largest fully connected part of a network (Ma, 2003). From the results, G4 has a smaller diameter of 20 and APL of 5.982 in comparison to M4 with diameter 31 and APL 7.076. Results from Broder *et al.* (2000) has shown that the diameter of the central core is at least 28, while the average length is 16. The difference in our results is because our network has considered too few nodes. Nevertheless, our results do imply that G4 has a stronger core and less divided clustered as compare to M4. This suggests that most of the websites crawled by the walkers from *google.com* has stronger connection with each other.

According to Figure 4 and 5, we notice that there is a rather straight line between a degree $1 < k < 6$ in all the networks. This shows that there is right-skewed degree distribution on all the networks where most nodes are of low degree. From the dataset, it is found that more 50% of the nodes are of degree one, meaning that the RW just passed through them only once. This might imply that more than 50% of the websites are not popular site where they lack the hyperlinks pointing to the site. As for those who have high degree, they are most probably famous website. Usually for famous website, it will have many hyperlinks pointing towards that website. Hence, it becomes a hub and the probability of RW to come back to the same site is much higher. Overall, networks from both the global and local websites show a similar pattern which is having power-law behaviour.

Table 2. Exponent of the power-law degree distribution and deviation values for the constructed networks

Network	In Degree	Out Degree
G1	(3.04±0.53)	(3.29±0.63)
G2	(2.77±0.26)	(2.91±0.28)
G3	(2.45±0.24)	(2.46±0.26)
G4	(2.20±0.20)	(2.47±0.22)
M1	(2.55±0.38)	(2.47±0.38)
M2	(2.72±0.41)	(3.41±0.51)
M3	(2.58±0.25)	(2.60±0.25)
M4	(2.54±0.21)	(2.70±0.23)

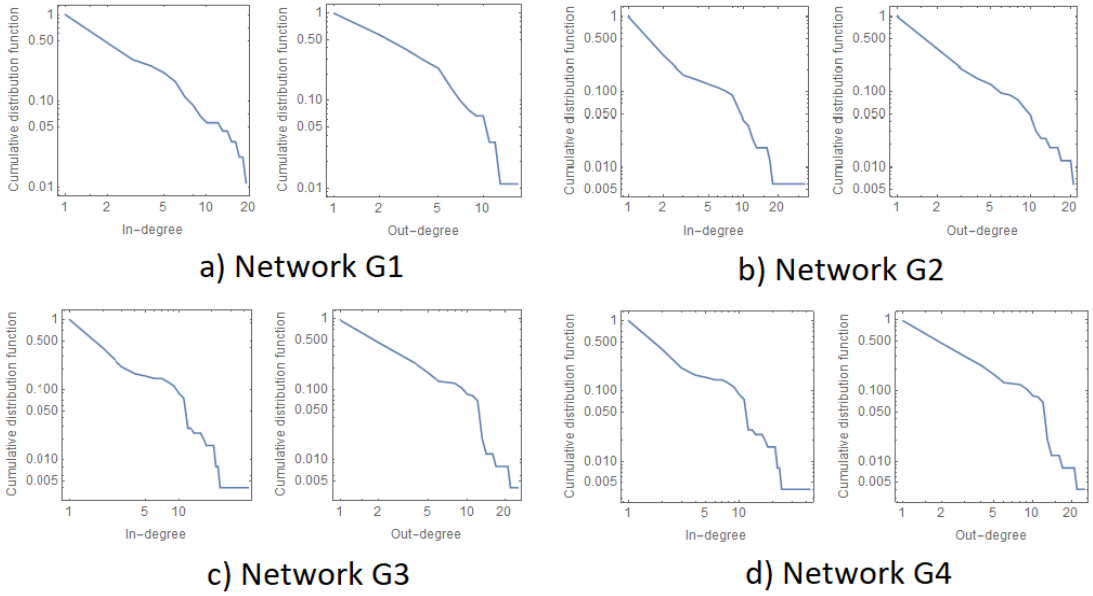

Figure 4. Cumulative in-degree and out-degree distribution for networks constructed from *google.com*

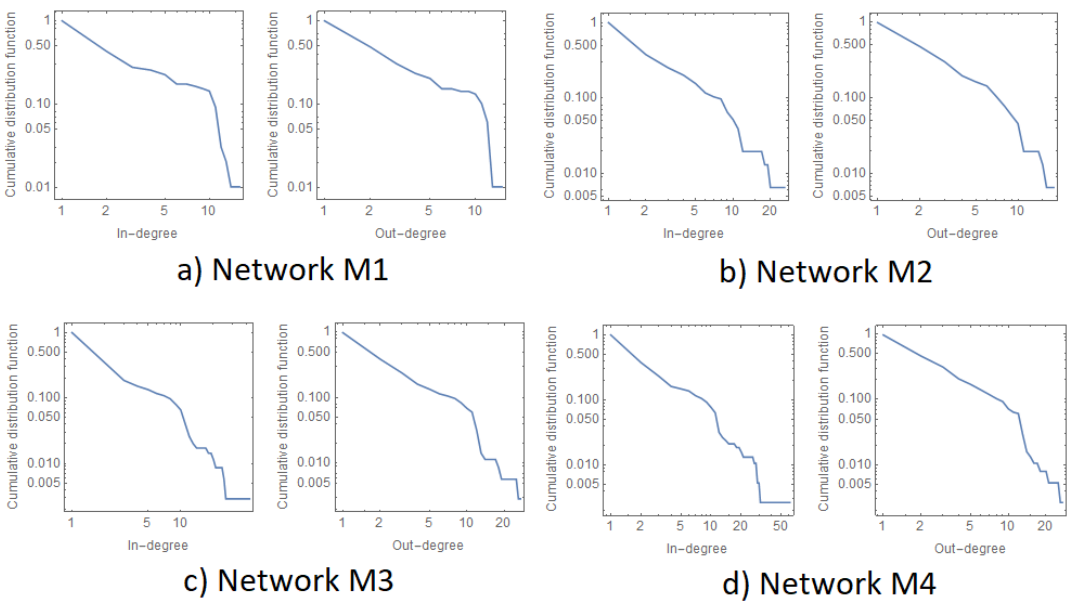

Figure 5. Cumulative in-degree and out-degree distribution for networks constructed from *mudah.my*

C. Distribution of Top-Level Domain (TLD)

From the nodes of the networks, we can further split the hostname into a list of components, e.g. *www.google.com* into {*www, google, com*}. Taking the last part of the return list from the programme, we manage to identify the top-level domain (TLD) or public suffixes from our networks. Five top-level domains for each network are presented in Table 3. The values in brackets represent the total number of TLDs tallied. The values at each TLDs represent the relative frequency of each TLD in a network. Websites from the top five TLDs have amounted to more than 80% of the total number of nodes of the network. TLD "*.com*" appear as top public suffix as it alone has covered more than 60% for all the networks. This is followed by "*.org*" ranging from 4% to 26%. This distribution is generally agreed with the report from Verisign Domain Industry Brief 2018 which mentioned "*.com*" is the largest domain registered with 135.6 million while domain "*.org*" comes in seventh place with 10.3 million. Total registered domain as of Jun 2018 is 339.8 million (*www.verisign.com*).

Country-code TLD (ccTLD) such as "*.my*", "*.eu*" and "*.ca*" only cover small fractions in the networks because their market fraction is very small in comparison to other TLDs. One reason for these small values is that TLD such as "*.com*" is a global domain which is used across all countries while ccTLD usually used in a specified country. According to the Malaysia Network Information Centre (MYNIC), domain "*.my*" only has 338185 active sites. By comparing all the networks from the global and local website, we found that the higher the number of random walkers, the higher number of domains are covered by the dynamical process. Also, we can see that domain such as "*.google*" appeared in top 5 for all networks sampled from *google.com* while "*.my*" appeared in all networks sampled from *mudah.my* except M3 which is actually at sixth place. These results give a clue where the network is sampled.

D. Centrality Analysis

Results for DC and PR measurement for network G4 and M4 are shown in Table 4. We list down the top 20 of the websites based on DC and PR values. From DC measurement, the most important website in G4 is *api.stackexchange.com* while in M4 is *appsource.microsoft.com*. For PR measurement, the most influential website in G4 is *cocatalog.loc.gov,* while *books.google.com* is most influential in M4. Comparing both networks, it is found that *cocatalog.loc.gov* appeared in all measurements. This site is part of the United State Copyright Office that maintains records for copyright registration in the United States. This shows that this website is quite popular

and influential as many sites linked it to them. Besides, from the results, we can see that domain *stackexchange.com* which is a network for question and answer websites for various fields appeared few times in both measurements. This implies that this domain is of great importance and popular among people in various fields such as physics, video games and patents.

Table 3: Five highest distribution for TLDs for global and local networks. The values in bracket show the total number of TLDs

G1 (12)	G2 (17)	G3 (17)	G4 (27)
com,0.79	com, 0.60	com, 0.68	com, 0.67
org, 0.07	org, 0.26	org, 0.10	org, 0.14
google, 0.03	gov, 0.04	gov, 0.06	gov, 0.05
my, 0.02	edu, 0.02	io, 0.04	uk, 0.02
co, 0.01	google, 0.01	google, 0.02	google, 0.01
M1 (9)	M2 (13)	M3 (19)	M4 (20)
com, 0.78	com, 0.77	com, 0.74	com, 0.67
org, 0.09	ca, 0.04	org, 0.07	org, 0.13
google, 0.04	org, 0.04	edu, 0.04	gov,0.04
io, 0.02	my, 0.03	gov, 0.03	edu, 0.03
my, 0.02	eu, 0.03	eu, 0.03	my, 0.02

Table 4: Top 20 websites ordered by degree centrality (DC) and PageRank (PR) for networks G4 and M4

	G4				M4			
	Website	DC	Website	PR	Website	DC	Website	PR
1	api.stackexchange.com	104	cocatalog.loc.gov	0.074302	appsource.microsoft.com	86	books.google.com	0.0536232
2	area51.stackexchange.com	76	cloudplatformonline.com	0.039389	arduino.stackexchange.com	57	appsource.microsoft.com	0.0435058
3	archive.is	42	chrome.google.com	0.037919	books.google.com	40	archive.org	0.0318027
4	app.unionmetrics.com	42	chemistry.stackexchange.com	0.034431	archive.org	37	arduino.stackexchange.com	0.0313294
5	about.ask.fm	36	cloud.google.com	0.034372	about.twitter.com	34	appstoreconnect.apple.com	0.0264024
6	cocatalog.loc.gov	35	coachella.tumblr.com	0.034372	analytics.facebook.com	33	cloudplatformonline.com	0.0244176
7	analytics.twitter.com	31	ca.wikipedia.org	0.034372	appstoreconnect.apple.com	31	curia.europa.eu	0.0226735
8	about.twitter.com	31	chromium.googlesource.com	0.034372	apple.stackexchange.com	28	cocatalog.loc.gov	0.0226735
9	en.wikipedia.org	29	channel9.msdn.com	0.034372	cars.booking.com	26	circleci.com	0.0226735
10	beinternetawesome.withgoogle.com	27	chemrxiv.org	0.031783	accounts.justia.com	26	copyright.gov	0.0225393
11	ar-ar.facebook.com	27	api.stackexchange.com	0.031197	chemistry.stackexchange.com	25	connect.telenordigital.com	0.0225393
12	chrome.google.com	25	chroniclingamerica.loc.gov	0.03085	ar-ar.facebook.com	25	com.com	0.0225393
13	anime.stackexchange.com	25	codegolf.stackexchange.com	0.03085	curia.europa.eu	24	code.msdn.microsoft.com	0.0225393
14	coachella.tumblr.com	23	circleci.com	0.029096	cocatalog.loc.gov	24	com.xyz	0.0225393
15	cloudplatformonline.com	23	area51.stackexchange.com	0.026488	circleci.com	24	commons.healthymaterials.net	0.0225393
16	cloud.google.com	23	archive.is	0.020166	ec.europa.eu	23	cloud.google.com	0.0225393
17	chromium.googlesource.com	23	app.unionmetrics.com	0.019356	copyright.gov	23	apple.stackexchange.com	0.0210726
18	channel9.msdn.com	23	bitbucket.org	0.017253	connect.telenordigital.com	23	accounts.justia.com	0.0191978
19	ca.wikipedia.org	23	biology.stackexchange.com	0.017234	com.xyz	23	company.justia.com	0.0184144
20	ads.twitter.com	23	en.wikipedia.org	0.012931	commons.healthymaterials.net	23	ar-ar.facebook.com	0.0152819

IV. CONCLUSION

In this work, we have visualised the aggregated version of the networks sampled from the global and local websites, namely *google.com* and *mudah.my* using random walk process. From the study, generally, all networks follow the common power-law degree distribution except for the networks G1 and M2. This is likely due to a number of vertices with higher degree, and also the tail does not go far as N is not big enough.

Having fixed the number of steps and varying the number of walkers has shown to impact the properties of the networks. Increasing of number of RW has shown to increase the number of edges and nodes, hence increased the coverage of the steps as well as the number of TLDs. In contrast, the network density decreases as the number of nodes increase,

forming a sparsely connected network.

Measurement of the connectivity of the network has suggests that most websites traversed by the walkers from the global site have a stronger connection with each other. From crawling time perspective, RW at all global site networks takes less time to complete the task in comparison to the local site. Based on domain analysis, main TLD of these networks is "*.com*" covering more than 60% of the total nodes followed by "*.org*" covering 4% to 26%. The appearance of ccTLD at local sites networks give a strong indication of the origin of the sampled networks.

From centrality analysis for network G4 and M4, *cocatalog.loc.gov* is the only site that appears in all measurements. This shows that this website is quite popular and influential as many sites in both networks linked it to them. Another important finding is that *stackexchange.com* is a popular domain as it appears a few times in PR and DC measurements for both networks.

revealed that the average diameter and path length from the global site is shorter than the one from the local site. This

V. ACKNOWLEDGEMENT

This work was supported by Universiti Putra Malaysia, under UPM Putra Grant with project number GP-IPS/2018/9624900.

I. REFERENCES

Aggarwal, A., Floudas, C.A. 1990, 'Synthesis of general distillation sequences-nonsharp separations', *Comput Chem Eng*, vol. 14, no. 6, pp. 631-653.

Albert, R., Jeong, H. & Barabási, A.L. 1999, 'Diameter of the World-Wide Web', Nature 401: 130.

Bar-Yossef, Z., Berg, A., Chien, S., Fakcharoenphol, J. & Weitz, D. 2000, 'Approximating Aggregate Queries about Web Pages via Random Walks', in *Proceedings of the 26th International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers Inc., pp. 535-544.

Barabási, A.L., Albert, R. & Jeong, H. 2000, 'Scale-free characteristics of random networks: the topology of the world-wide web', *Physica A: Statistical Mechanics and its Applications*, vol. 281, no. 1, pp. 69-77.

Broder, Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener J. 2000, 'Graph structure in the Web', *Computer Networks*, vol. 33, no. 1, pp. 309-320.

Clauset, A., Shalizi, C.R. & Newman, M.E.J. 2009, 'Power-Law Distributions in Empirical Data', *SIAM Review*, vol. 51, no. 4, pp. 661-703.

Csermely, P., London, A., Wu, L.Y. & Uzzi B. 2013, 'Structure and dynamics of core/periphery networks', *Journal of Complex Networks*, vol. 1, no. 2, pp. 93-123.

Hall, W. & Tiropanis, T. 2012, 'Web evolution and Web Science', *Computer Networks*, vol. 56, no. 18, pp. 3859-3865.

Lehmberg, O., Muesel, R. & Bizer, C. 2014, 'Graph structure in the web: aggregated by pay-level domain', in *Proceedings of the 2014 ACM conference on Web science*, Bloomington, Indiana, USA, ACM, pp. 119-128.

Ma, Hong-Wu & Zeng, An-Ping 2003, 'The Connectivity Structure, Giant Strong Component and Centrality of Metabolic Networks', *Bioinformatics*, vol. 19, no. 11, pp. 1423–1430.

DOI: 10.1093/bioinformatics/btg177.

Meusel, R., Vigna, S., Lehmberg, O. & Bizer, C. 2015, 'The Graph Structure in the Web – Analyzed on Different Aggregation Levels', *The Journal of Web Science*, vol. 1, no. 1, pp. 33-47.

Newman, M. 2010, *Networks: An Introduction*, Oxford University Press, Inc., UK.

Rusmevichientong, P., Pennock, D.M., Lawrence, S. & Giles, L. 2001, *Methods for Sampling Pages Uniformly from the World Wide Web*, AAAI Technical Report FS-01-04.

Silvestri, T. 2013, 'Random Walks on the World Wide Web', *The Mathematica Journal*, vol. 15.