Extracting Domain Knowledge by Complex Networks Analysis of Wikipedia Entries

Neven Matas*, Sanda Martinčić-Ipšić**, Ana Meštrović**
*Faculty of Humanities and Social Sciences, University of Rijeka, Croatia
**Department of Informatics, University of Rijeka, Croatia
nmatas@ffri.hr, smarti@uniri.hr, amestrovic@inf.uniri.hr

Abstract - In this paper we describe a complex networks analysis of Wikipedia. We construct 10 different networks from Wikipedia entries (articles) related to the chosen domain. The goal of the experiment is to extract domain knowledge in terms of identifying entries that are centrally positioned and entries that are strongly related. We apply complex networks analysis on all acquired networks and examine the networks' structure. We employ centrality measures in order to find centrally positioned entries in the network. Furthermore we identify communities and find which entries are densely connected according to the network structure.

I. INTRODUCTION

Complex networks exhibit specific topological features, such as high clustering coefficients, small diameters, a power-law degree distribution, community structure, one or several giant components, hierarchical structures, etc. Two important classes of complex networks that can be further differentiated are small-world networks [1] with small distances and high clustering coefficients as main properties and scale-free networks [1] which can be characterized by a power-law degree distribution.

Wikipedia can be modelled as a complex network in a way that Wikipedia entries are nodes, and links between two nodes are established if there is a hyperlink between these two entries. Early attempts to quantify Wikipedia using complex networks analysis were focused only on network structure of linked Wikipedia entries. In [2] Zlatić et al. present an analysis of Wikipedias in several languages as complex networks. They show that many network characteristics (degree distributions, growth, topology, reciprocity, clustering, assortativity, path lengths and triad significance profiles) are common to Wikipedias in different languages and show the existence of a unique growth process. The same authors studied Wikipedia growth based on information exchange in [3]. In [4] an analysis of the statistical properties and growth of Wikipedia is presented. Pemble and Bingol [5] have constructed two complex networks out of English and German Wikipedia corpora and analyzed conceptual networks in different languages.

This work has been supported in part by the University of Rijeka under the project number 13.13.2.2.07.

The other research direction is focused on content found on Wikipedia and analyses Wikipedia as a (domain) knowledge network. In Fang [6] they first extract a specific domain knowledge network from Wikipedia (specifically, four domain networks on mathematics, physics, biology, and chemistry) and then carry out statistical analysis on these four knowledge networks. Also, they show that MathWorld and Wikipedia Math share a similar internal structure. In [7] Masucci et al. extract the topology of the semantic space and measure the semantic flow between different Wikipedia entries. They further analyze a directed complex network of semantic flow. In [8] the results of semantic language networks analysis are presented in general.

Motivated by the second approach that studies Wikipedia as a knowledge network, we wanted to study how the network structure is related to domain knowledge. The goal of our experiment was to extract centrally positioned entries in the network and analyze how these entries are related to domain knowledge and are some more important than other. In the second part of the experiment the task was to extract entries that belong to the same community and check whether they are semantically related.

In our previous research, we have already analyzed Wikipedia as a complex network [9], but by constructing a network of syllables. Also, we examined the structure of Croatian language networks in [10,11,12]. In [12,13] we applied network measures for a keyword extraction task. In all our previous experiments we were focused solely on language structure and this is our first attempt to analyze semantic relations in a network.

In the second section we present key measures of complex networks involved in network structure analysis. In the third section we describe data sources and network construction principles. In the fourth section we present the results. Finally, the fifth section contains a conclusion and possible directions for future research.

II. NETWORK STRUCTURE ANALYSIS

In this section we review some of the most important network measures [14]. Every network has an N number of nodes and a K number of links. The degree

of a node *i* is the number of links with which the node is connected, k_i. Considering the fact that we are working with directed networks, we must specify two types of degrees: the in-degree, k_i^{in}, corresponding to the number of incoming links and the out-degree, k_i^{out}, equal to the number of outgoing links for any particular node *i*. The average degree of the network is:

$$< k > = \frac{2K}{N}. \tag{1}$$

For the directed networks we omit multiplication by 2. In the further equations we assume that the network is directed and that the total possible number of links is equal to $N(N\text{-}1)$. For every two connected nodes *i* and *j*, the number of connections lying on the path between them is represented as d_{ij}, and so d_i is the average distance of a node *i* from all other nodes, and it is obtained by:

$$d_i = \frac{\sum_j d_{ij}}{N}. \tag{2}$$

For the next two measures, if a network contains more than one component, we consider the largest component. The average shortest path length between every two nodes in a network is:

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}. \tag{3}$$

And the maximum distance results in the network diameter, *D*:

$$D = max_i d_i. \tag{4}$$

The clustering coefficient is a measure which defines the presence of connections between the nearest neighbours of a node. And so, c_i (the clustering coefficient) of a node is a fraction between the number of edges E_i that exist between that k_i and the total possible number:

$$c_i = \frac{2E_i}{k(k-1)}. \tag{5}$$

The average clustering coefficient of a network is defined as the average value of the clustering coefficients of all nodes in a network:

$$C = \frac{1}{N} \sum_i c_i. \tag{6}$$

Density of a network is a measure of network cohesion defined as the number of observed links divided by the number of total possible links:

$$d = \frac{K}{N(N-1)}. \tag{7}$$

Degree centrality of a node *i* is the degree of that node. It can be normalised by dividing it by the maximum possible degree $N-1$:

$$dc_i = \frac{k_i}{N-1}. \tag{8}$$

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Let σ_{jk} be the number of shortest paths from node *j* to node *k* and let $\sigma_{jk}(i)$ be the number of those paths that pass through the node *i*. The normalised betweenness centrality of a node *i* is given by:

$$bc_i = \frac{\sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}}{(N-1)(N-2)}. \tag{9}$$

Closeness centrality is defined as the inverse of farness, i.e. the sum of the shortest distances between a node and all other nodes. Let d_{ij} be the shortest path between nodes *i* and *j*. The normalised closeness centrality of a node *i* is given by:

$$cc_i = \frac{N-1}{\sum_{i \neq j} d_{ij}}. \tag{10}$$

Modularity measures the quality of the network partition in the communities. The modularity of a network partition is a scalar value between -1 and 1 that measures the density of links inside communities as compared to links between communities. Let e_{ij} be the fraction of edges in the network that connect vertices in group *i* to those in group *j*, and let $a_i = \sum_j e_{ij}$. Then the modularity can be calculated using following equation:

$$Q = \sum_{i=1}^{k} \left(e_{ii} - a_i^2 \right). \tag{11}$$

The degree assortativity coefficient measures the tendency of nodes in a network to connect to nodes similar to themselves. The coefficient lies between -1 and 1 and it is quantified via the Pearson correlation. Positive *r* values indicate a correlation between similar-degree nodes. Let q_k and q_j be the distribution of the degree of out-edges that do not connect to the other node in question, e_{jk} the joint probability distribution of q_k and q_j, and σ_q^2 the variance of the distribution. Then we can calculate the assortativity coefficient using the following equation:

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}. \tag{12}$$

On the meso-scale level complex networks analysis includes a community detection task [15]. Communities, in this sense, are groupings of densely interconnected nodes within a network. In other words,

nodes in a community have a greater amount of connections amongst each other than with other nodes in the network. Several algorithms are used for community detection such as hierarchical clustering, Girvan-Newman's algorithm, minimum-cut method, etc. One of the most efficient is the Louvain method [16], a greedy optimization method that optimizes the modularity of a network's partitions. The number of communities (*Nc*) represents the amount of such groupings found within a network.

III. Network Construction

For the purpose of our experiment we collect entries from Wikipedia and construct networks related to the domain. Our intention was to construct two types of networks: level 2 networks and level 4 networks. We construct level 2 networks by starting with a chosen seed entry (e.g. "*Complex network*" or "*Data*"), storing all the hyperlinks to related entries from the seed entry's text (level 1) and proceeding to extract the hyperlinks from all the entry pages taken from the original entry (level 2). Analogously, we construct level 4 networks by taking the first 10 hyperlinks from a given entry page and proceeding to repeat the task three times, arriving at level 4. We limit the hyperlinks to the first 10 due to the computational complexity at the same time having in mind that the most general hyperlinks are usually at the beginning of the entry's text.

Therefore, the first task is the construction of a web scraping program which would extract hyperlinks from a Wikipedia entry's text. The hyperlinks are extracted using a Python package for HTML parsing called Beautiful Soup which parses the HTML structure of a given HTML document into a parse tree. By navigating the tree we locate the tag ID which corresponds to article content ("mw-content-text") and proceed to extract the hyperlinks which themselves are found within paragraph (<p>) tags and finally inside link (<a>) tags in that section of the page. Finally, each network is stored in an edge list in the following format: "entry title" \t "linked entry title". We had some difficulties with processing non-ASCII script and hyperlinks that weren't connected to other documents (citations, in-page references, etc.), but we managed to avoid those by checking the data during the extraction process.

In our directed network, each entry's title represents a node and it is connected to other entries hyperlinked in its text, again represented as network nodes. We construct a total of 10 domain networks for five chosen seed entries: "*Byte*", "*Complex network*", "*Computer science*", "*Data*" and "*Programming language*". The naming scheme includes the level of a specific network in its name (e.g. the level 2 network for "Byte" is BT2). Since we consider unweighted networks, we dismiss double links. This, along with the fact that some entries do not contain 10 hyperlinks resulted in our level 4 networks having less than 10^4 expected edges. We use Python and the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [17].

IV. Results

In this section we present the results of our measuring described in section 2, such as average degree $\langle k \rangle$, average path distance L, diameter D, average clustering coefficient C, density d, modularity Q, number of communities (N_c) and degree assortativity coefficient r. We also present the most central nodes (according to the three centrality measures) and communities in networks detected by using the Luvain algorithm.

In Table I. we present estimated global network measures. There are certain differences between measures for level 2 and 4 which are evident upon closer inspection. For instance, level 4 networks have significantly larger average path lengths, diameters, assortativity coefficients, often a significantly larger number of detected communities and slightly larger average degrees. The modularity measure and density are comparable between the two, whilst level 2 networks show larger clustering coefficients.

For comparison with random networks, the table also includes two measures for equivalent random networks (Erdös-Renyi random graphs) – the average shortest path length ($L_{ER} = \ln N / \ln \langle k \rangle$) and the average clustering coefficient ($C_{ER} = \langle k \rangle / N$). The results show that the complex networks we have constructed have a significantly higher average clustering coefficient than their Erdös-Renyi random graph counterparts. This, in addition with a relatively small average shortest path length L led us to conclude that we are dealing with small-world networks as described by Watts and Strogatz [20]. For the purposes of this comparison we treat the networks as undirected.

Moreover, a distinctly high modularity coefficient Q (higher than 0.7 in all but one network, as visible in Table I.) shows a clear tendency towards community clustering of nodes present in the networks. We did not observe any strict rule governing community size across networks, although level 2 networks have an understandably smaller N_c which we contributed to the very construction principle as described in section 3.

Figure 1 and Figure 2 show the degree distribution plots for level 2 and level 4 networks corresponding to the "computer science" and "complex network" entries respectively. The observed difference between the level 2 and 4 networks is believed to be due to the method of network construction.

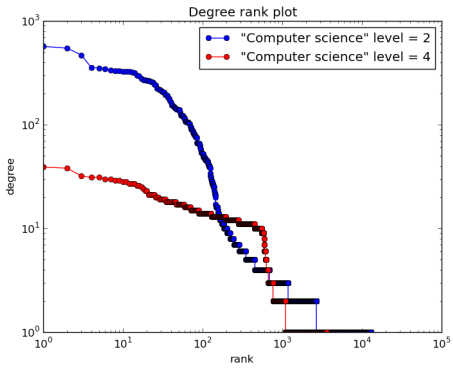

Figure 1 - Degree distribution for CS2 & CS4

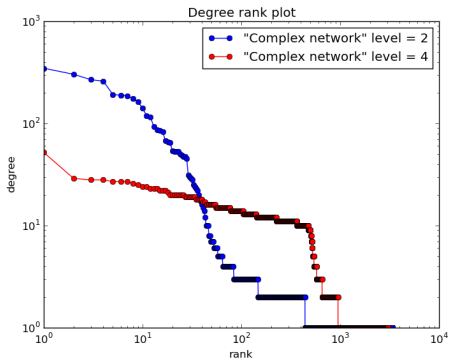

Figure 2 - Degree distribution for CN2 & CN4

TABLE I. GLOBAL NETWORK MEASURES CALCULATED FOR ALL 10 NETWORKS

Measure	"Byte"		"Complex network"		"Computer science"		"Data"		"Programming Language"	
Network	BT2	BT4	CN2	CN4	CS2	CS4	DT2	DT4	PL2	PL4
Number of nodes (N)	3945	3632	3405	3070	12881	3630	2297	3658	7467	3965
Number of edges (K)	5112	5611	4132	5008	18852	5851	2630	5531	13933	6215
Average degree ($<k>$)	1.296	1.545	1.214	1.631	1.464	1.612	1.145	1.512	1.145	1.612
Avg. shortest path (L)	3.693	6.834	3.198	9.218	3.417	6.277	3.086	6.369	3.127	6.277
Avg. shortest path (L_{ER})	8.693767	7.2662195	9.168408	6.791134	8.8088393	7.0022451	9.340827	7.4144377	10.763658	7.0776521
Diameter (D)	9	15	6	22	7	14	7	14	6	22
Average clustering coefficient (C)	0.06	0.021	0.043	0.024	0.074	0.019	0.043	0.019	0.082	0.021
Average clustering coefficient (C_{ER})	0.000657	0.0008508	0.0007131	0.0010625	0.0002273	0.0008882	0.000997	0.0008267	0.0003067	0.0008131
Density (d)	0.0003	0.00042	0.00035	0.00053	0.00011	0.00044	0.00049	0.00041	0.00025	0.0004
Modularity (Q)	0.778	0.776	0.794	0.763	0.725	0.771	0.828	0.779	0.594	0.78
Number of communities (N_c)	17	32	17	21	23	27	18	31	19	30
Degree assortativity coefficient (r)	-0.592	-0.048	-0.521	0.021	-0.491	-0.028	-0.561	-0.048	-0.468	-0.059

After the analysis on the global level, we analyse the networks on the local level in terms of centrality measures. Tables II. and III. show lists of top ten entries according to the three centrality measures for the two seed entries: "*Computer science*" and "*Programming language*". We analyse the degree centrality, betweenness centrality and closeness centrality. For the degree centrality we treated the network as undirected. For each centrality measure and domain there are two lists of entries, one for level 2 networks and another for level 4 networks. We noticed that the lists for level 2 networks consist of entries that are semantically related to the seed entries ("*Computer science*" or "*Programming language*") in a way that might be ascribed as belonging to a hierarchy. This is especially evident for the closeness centrality measure. For example, the list of top ten entries according to the closeness centrality for the seed entry "*Computer science*" contains other scientific domains (theoretical computer science, mathematics, artificial intelligence, physics, engineering) and for the seed entry "*Programming language*", list contains some prominent programming languages (C, Java, Perl, Python, C++).

TABLE II. TOP TEN ENTRIES IN THE „*COMPUTER SCIENCE*" NETWORKS (CS2, CS4) REGARDING THE THREE CENTRALITY MEASURES: DEGREE CENTRALITY, BETWEENNESS CENTRALITY AND CLOSENESS CENTRALITY

	Degree centrality		Betweenness centrality		Closeness centrality	
	CS2	CS4	CS2	CS4	CS2	CS4
#1	*human*	*mathematics*	***computer science***	*computer science*	***computer science***	*computer science*
#2	*university of cambridge*	*cell (biology)*	*computer*	*information*	***mathematics***	*information*

#3	***philosophy***	*computer science*	***mathematics***	*protein*	***theoretical computer science***	*science*
#4	*industrial revolution*	*computer*	***artificial intelligence***	*science*	*computer*	*cell (biology)*
#5	*gottfried wilhelm leibniz*	*information*	***philosophy***	*algorithm*	***artificial intellgience***	*mathematics*
#6	***physics***	*protein*	*human*	*logic*	***philosopy***	*ancient greek*
#7	***eletrical engineering***	*organism*	*gottfried wilhelm leibniz*	*organism*	***physics***	*latin*
#8	***artificial intelligence***	*dna*	*algorithm*	*cell (biology)*	*human*	*computing*
#9	***mathematics***	*computer program*	***theoretical computer science***	*computing*	*gottfried wilhelm leibniz*	*algorithm*
#10	*alan turing*	*philosophy*	***physics***	*mathematics*	*engineering*	*bit*

TABLE III. TOP TEN ENTRIES FOR THE „*PROGRAMMING LANGUAGE*" NETWORKS (PL2, PL4) REGARDING THE THREE CENTRALITY MEASURES: DEGREE CENTRALITY, BETWEENNESS CENTRALITY AND CLOSENESS CENTRALITY

	Degree centrality		Betweenness centrality		Closeness centrality	
	PL2	PL4	PL2	PL4	PL2	PL4
#1	*history of computing hardware*	*mathematics*	*programming language*	*programming language*	*programming language*	*programming language*
#2	*internet*	*computer*	*computer*	*computer*	***c (programming language)***	*ancient greek*
#3	*english language*	*computer science*	***c (programming language)***	*software engineering*	*computer programming*	*computer*
#4	*computer*	*physics*	*compiler*	*computing*	***java (programming language)***	*mathematics*
#5	***c (programming language)***	*set (mathematics)*	*english language*	*computer science*	***perl***	*arithmetic*
#6	***python (programming language)***	*greek language*	*computer program*	*algorithm*	*compiler*	*science*
#7	*university of manchester*	*logic*	*internet*	*message*	*computer program*	*greek language*
#8	***perl***	*language*	***perl***	*communication*	***python (programming language)***	*physics*
#9	***programming language***	*central processing unit*	***python (programming language)***	*machine*	*control flow*	*latin*
#10	***php***	*electronics*	***java (programming language)***	*function (mathematics)*	***c++***	*computer science*

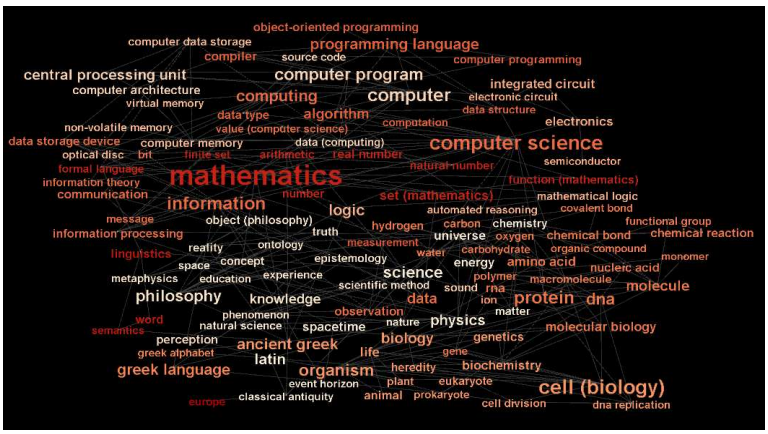

Figure 3 – Communities in the CS2 network

In the second part of the experiment we analyse communities in all 10 networks in order to explore which entries are grouped together. Figure 3 shows most significant entries from the CS2 network grouped into communities. Different communities are presented in different colours. For example, entries related to the mathematics domain (*mathematics*, *number*, *set*, *function*, *real number*, etc.) are in the red-coloured community; entries related to the computer science domain (*computing*, *algorithm*, *compiler*, etc.) are in the orange-coloured community; entries that are related to the biology domain (*cell*, *organism*, *gene*, etc.) are in the light-orange coloured community and entries that are related to the philosophy domain (*reality*, *concept*, *knowledge*, etc.) are in the white-coloured community. It can be observed that entries grouped into communities are more closely semantically related than entries from different communities. The results are similar for other networks; semantically related entries are grouped into communities much more than entries that are not semantically related.

V. Conclusion

In this paper we present our initial attempt to study Wikipedia as a complex network. We extract parts of Wikipedia related to 5 chosen seed entries. We construct 10 different networks using two different principles of construction. Then we analyse the global structure of all networks. We show that all networks have similar properties: a high average clustering coefficient in comparison to the random networks, small distances, low density and community structure. From these global measures we may conclude that all 10 networks extracted form Wikipedia are small-world networks. These results are in line with previous studies of Wikipedia as a complex network.

Furthermore, we explore semantic relations in the constructed networks. We use network centrality measures to extract entries in the networks that are significant according to the network structure. Three centrality measures are employed for this task: degree centrality, betweenness centrality and closeness centrality. It can be observed that for level 2 networks centrality measures obtain good results (especially closeness centrality). Among top ten entries according to the closeness centrality are entries that are semantically related to the domain. This can be useful for modelling taxonomy or domain ontology. Furthermore, semantically related entries are grouped into communities more often that entries that are not semantically related.

These findings can be partially explained as a consequence of network construction rules employed in this experiment. However, these preliminary results suggest that Wikipedia is well organised and its structure can be captured and explored by a complex networks approach. In future work we plan to extract a broader section of Wikipedia and explore its potential as a knowledge network. We will study the domain knowledge extraction possibilities and perform the evaluation of the results.

VI. References

[1] A. L., Barabasi and R. Albert, "Emergence of scaling in random networks", Science, 286, pp. 509–512, 1999.

[2] V. Zlatić et al., "Wikipedias: Collaborative web-based encyclopedias as complex networks," in Physical Review E 74.1 (2006): 016115.

[3] V. Zlatić and H. Štefančić, "Model of wikipedia growth based on information exchange via reciprocal arcs," EPL (Europhysics Letters) 93.5 (2011): 58005.

[4] G. Caldarelli, A. Capocci, V. Servedio, L. Buriol, D. Donato, and S. Leonardi, "Preferential attachment in the growth of social networks: the case of Wikipedia," in APS Meeting Abstracts, vol. 1, p. 33003. 2006.

[5] F. C. Pemble and H. Bingol, "Complex Networks in Different Languages: A Study of an Emergent Multilingual Encyclopedia", Unifying Themes in Complex Systems, 1, 612, 2008.

[6] Z. Fang, J. Wang, B. Liu, W. Gong, "Wikipedia as domain knowledge networks: domain extraction and statistical measurement," in Proceedings of international conference on knowledge discovery and information retrieval (KDIR 2011), Paris, Oct 2011

[7] A.P. Masucci, A. Kalampokis, V. Martínez Eguíluz, and E. Hernández-García. (2011) "Wikipedia Information Flow Analysis Reveals the Scale-Free Architecture of the Semantic Space," PLoS ONE 6(2): e17333. doi:10.1371/journal.pone.0017333

[8] J. Borge-Hothoefer and A. Arenas, "Semantic Networks: Structure and Dynamics," in Entropy, 12, pp. 1264-1302, 2010.

[9] K. Ban, I. Ivakić, and A. Meštrović, "A preliminary study of Croatian language syllable networks." In Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th, pp. 1296-1300. IEEE, 2013.

[10] D. Margan, S. Martincic-Ipšic, and Ana Meštrovic, "Network Differences Between Normal and Shuffled Texts: Case of Croatian," in Complex Networks V, pp. 275-283. Springer International Publishing, 2014.

[11] D. Margan, A. Meštrović and S. Martinčić-Ipšić, "Complex networks measures for differentiation between normal and shuffled Croatian texts," in Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on, pp. 1598-1602. IEEE, 2014.

[12] S. Šišović, S. Martinčić-Ipšić, and A. Meštrović, "Toward Network-based Keyword Extraction from Multitopic Web Documents," in Proceedings of 6th International Conference on Information Technologies and Information Society (ITIS2014), Šmarješke toplice, pages 18-27, Slovenia, 2014.

[13] S. Beliga, A. Meštrović and S. Martinčić-Ipšić, "Toward Selectivity Based Keyword Extraction for Croatian News," in CEUR Proceedings (SDSW 2014), Vol. 1301, pages 1-14, Riva del Garda, Trentino, Italy, 2014.

[14] M. E. J. Newman. "Networks: An Introduction", Oxford University Press, 2010.

[15] S. Fortunato, "Community detection in graphs," in Physics Reports 486, no. 3 (2010): 75-174.

[16] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of Statistical Mechanics: Theory and Experiment 2008, no. 10 (2008): P10008.

[17] A. A. Hagberg, D. A. Schult and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, 2008.