Structural Analysis of Relevance Propagation Models

Supplementary Material

1. Theoretical definitions of Centrality and Connectivity Measures

The analysis of the connectivity patterns and centrality measures of ontology graphs allows to obtain a better understanding of object importance and relevance among objects, while helping to place important constraints on RPM's and semantic similarity. Next, we describe the connectivity and centrality measures used in this work, for a generic graph *G = (V, E)*, and considering the distance between nodes *i* and *j*, *d(i, j)*, as the shortest path length between these nodes.

- **Graph density**: This measure is the proportion between the edges actually present in a graph *G* and the number of possible links that could be established, and is computed as follows:

$$Density(G) = \frac{|E|}{|V|(|V|-1)}$$

- **Diameter**: The largest distance between any pair of connected nodes, considering distance as the length of the shortest path.

- **Characteristic (Average) Path Length (CPL)**: This is the mean length of all the shortest path lengths in the graph. The formula for computing the *CPL* l is:

$$l(G) = \frac{1}{|V|*(|V|-1)} \sum_{i \in V} \sum_{j \in V \setminus \{i\}} d(i,j)$$

- **Connectivity length (CL)**: An alternative to the *CPL* proposed in Marchiori & Latora (2000). This measure employs the harmonic mean instead of using the arithmetic mean of the shortest path lengths, attempting to address the problem of disconnected nodes. Supposing that the distance between unreachable nodes is ∞ and $\infty^{-1} = 0$, the *CL* is computed as:

$$D(G) = \frac{|V|(|V|-1)}{\sum_{i,j \in G} \frac{1}{d(i,j)}}$$

- **Local Clustering Coefficient**: This coefficient is a degree of interconnection between the neighbors of a node (Watts & Strogatz, 1998). For a given node *i* in a directed graph, each neighbor is another node directly connected to it by an edge. For the case of undirected graphs, the clustering coefficient quantifies how close the group of neighbors is to being a clique. The number of real edges between neighbors of the corresponding node *i* is divided by the total amount of possible edges between them, according to the next formula:

$$C_i = \frac{\left|\left\{e_{jk} : j,k \in N_i, e_{jk} \in E\right\}\right|}{|N_i|\left(|N_i|-1\right)}$$

  Where:
  - N_i is the set of neighbors of node *i*
  - *(j, k)* is an edge between nodes *j* and *k*
- **Betweenness Centrality**: This measure is the proportion of shortest paths between every pair of nodes in the graph that pass through node *i* (Freeman, 1977). An efficient algorithm for its computation has been developed by Brandes (2001). This measure is computed as follows:

$$bc_i = \sum_{u \neq i \neq v} \frac{\sigma_{uv}(i)}{\sigma_{uv}}$$

Where:

- σ_{uv} is the total number of shortest paths from a node *u* to another node *v*
- $\sigma_{uv}(i)$ is the number of shortest paths from node *u* to node *v* that pass through *i*

- **Closeness Centrality**: The closeness of a node *i* is defined as the inverse of the sum of the shortest distances to that node from every other node in the network (Bavelas, 1950). The distance between non-reachable nodes is considered as zero. The formula is given by:

$$cl(i) = \frac{1}{\sum_{d(j,i)<\infty} d(j,i)}$$

- **Harmonic Centrality**: Another way for evaluating centrality of a node *i* is to compute the harmonic mean over all distances from every node *j* to node *i*, for the set of co-reachable nodes of *i* (Marchiori & Latora, 2000). Then, this measure is calculated as follows:

$$hc(i) = \sum_{d(j,i)<\infty,\, j \neq i} \frac{1}{d(j,i)}$$

- **Lin's index for Closeness Centrality**: An alternative index of centrality was presented in Lin (1976). It weights closeness by the square of the number of co-reachable nodes. Thus, the definition of this measure is:

$$lin(i) = \frac{|\{j | d(j,i)<\infty\}|^2}{\sum_{d(j,i)<\infty,\, j \neq i} d(j,i)}$$

This index considers closeness not as the inverse of a sum of distances, but rather as the inverse of the average distance. This way, a normalization in the value of closeness is attained. The square in the formula corresponds to a second multiplication by the number of co-reachable nodes, for giving a greater value of centrality to those nodes with a greater set of co-reachable nodes.

2. Complex Network special phenomena

- **"Small World" Networks**

The Small-World property is present in many complex networks, indicating a particular topology. As stated in Watts & Strogatz (1998), features such as robustness, propagation speed, computational power and synchronizability are highly related to this topological aspect. Numerically, these networks exhibit a low *CPL* l*(G)* and high levels of clustering coefficient. These characteristics turn any node reachable in relatively few steps from any other node (low *CPL*), and result in high connectivity for the entire graph (high clustering coefficient). Small-World networks could be placed on an intermediate point between regular lattices and random graphs.

As the analyses of Watts and Strogatz require total connectedness, the existence of isolated groups of nodes turns measures such as the *CPL* unreliable in the characterization of the complete network behavior. Moreover, the weight of each edge is not included in their analyses. To overcome these issues, the work of Marchiori & Latora (2000) introduces the use of harmonic means instead of classical geometric means. This change allows a more

accurate study of networks that are not fully connected. Also, it allows to study both metrical networks (with valued edges) and their topological abstraction (without valued edges).

- **Power-law distributions**

Multiple natural phenomena exhibit probability distributions that are said to be Power-law and are modeled with the following probability distribution function (Newman, 2005):

$$p(x) = Cx^{-\alpha}$$

The negative exponent α expresses the scaling exponent of each analyzed power-law distribution. When such a value is greater than 2, the distribution is said to have a well-defined mean, and if it is higher than 3, the associated phenomenon also has a finite variance. Natural phenomena commonly exhibit power-law distributions with α values greater than 2 but lower than 3 (Newman, 2005). Networks that exhibit a power-law degree distribution are said to be scale-free.

- **Visual analysis**

While numerical measures are a rich source of information, visualizations are complementary for the analysis of the topological aspects of a network. In our study we use the Large Networks Visualization tool (LaNet-Vi, Alvarez-Hamelin et al, 2005) to generate visualizations of the undirected graphs associated with the analyzed models employing a k-core decomposition algorithm. The k-core decomposition (Seidman, 1983) systematically identifies layers of nodes with equal degree. Each group of nodes obtained after each iteration constitutes a shell, and its nodes are identified with a specific color. In the visualization, the position of a node is a function of its shell and the average degree of its neighbors. The outermost layer of a k-core decomposition is made up by the least connected nodes, and the central layer by the most connected ones.

In Figure SM-1 the topology of a network is graphically represented by means of the LaNet-Vi tool. The nodes are displayed over different circles, according to several criteria:

- Nodes included in the same circle belong to the same shell. Every node has a color associated with its shell.
- Nodes with the same size have equal degrees, and could be in different shells.
- The diameter of a shell is a function of the number of nodes within it.
- Shells represented by concentric circles are related to the same connected component.
- The position of each node inside the circle is computed as a function of its links.
- Besides the shell a node belongs to, the visualizations group nodes belonging to the same cluster together.

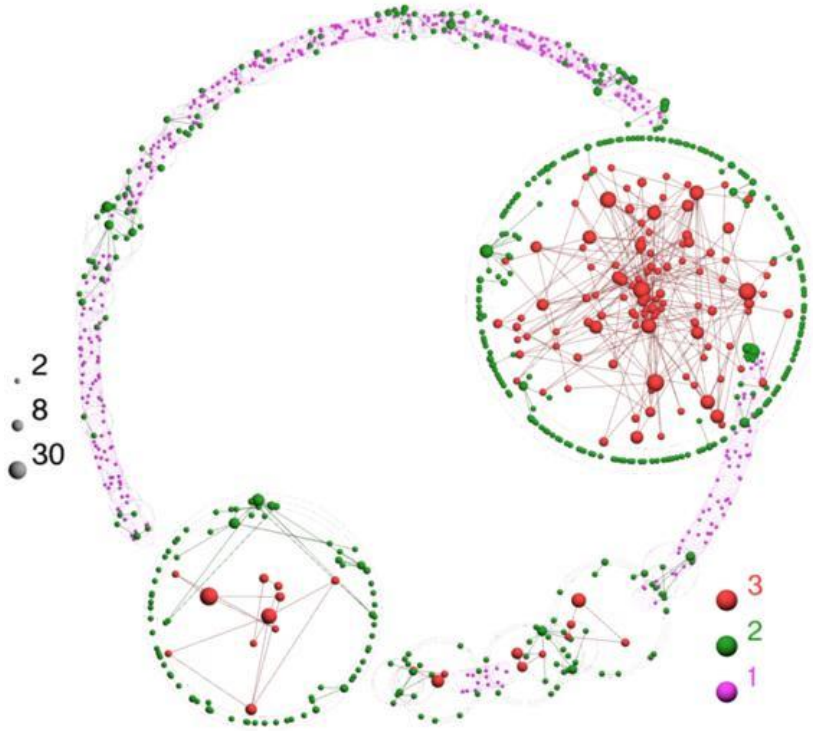

Figure SM-1: Graphical representation of a k-core decomposition.

3. Values of Centrality and Connectivity Measures for the DMOZ models

In this section we provide additional values of centrality and connectivity for the RPM's computed from the DMOZ graph, to allow further analyses and arguments. We show tables and figures that illustrate complementary facts about the studies shown in the main article. Table SM-1 shows the clustering coefficient values (average, number of nodes with CC=0 and number of nodes with CC=1) and Table SM-2 enumerates the scaling exponents associated with each RPM.

Table SM-1: Clustering coefficients values of DMOZ models

Model	Average CC	Number of nodes with CC=0	Number of nodes with CC=1
T	0	571148	0
S	0.0096	548,312	131
R	0.0484	496,157	2,040
G1	0.0531	372,272	935
G2	0.0754	372,272	4,103
G3	0.4823	0	0
G4	0.4078	0	0
G5	0.1092	363,764	323

G6	0.0897	496,157	41,273
G7	0.0179	489,025	30
G8	0.4261	0	0
G9	0.0548	464,308	8,881
G10	0.0810	464,308	22,280
G11	0.4652	0	0
G12	0.4660	0	0
G13	0.3975	0	0
G14	0.4352	0	0
G15	0.2420	0	0
G16	0.2785	0	0
G17	0.2793	0	0
G18	0.3956	0	0
G19	0.4406	0	0
G20	0.4064	0	0

Table SM-2: Scaling exponents of in-degree, out-degree and degree (for underlying undirected graphs) distributions for DMOZ models

Model	γ^{i}	γ^{o}	γ
T	-	-	-
S	3.215	2.03	-
R	2.012	3.873	-
G_1	2.242	2.265	2.339
G_2	2.302	2.221	2.339
G_3	-	0.911	0.958
G_4	1.987	0.997	1.047
G_5	1.434	0.936	-
G_6	2.121	2.121	-
G_7	2.652	2.220	-
G_8	2.372	0.980	-

G_9	2.168	2.109	-
G_{10}	2.225	2.101	-
G_{11}	1.990	0.989	-
G_{12}	2.080	1.028	-
G_{13}	1.543	0.917	-
G_{14}	2.134	1.063	1.155
G_{15}	1.934	0.727	0.929
G_{16}	1.351	0.594	-
G_{17}	1.346	0.594	0.825
G_{18}	2.221	0.924	1.042
G_{19}	2.181	0.944	1.099
G_{20}	2.017	1.023	-

Betweenness centrality

The Betweenness Centrality measure helps to identify influential nodes in a network. Such an assessment of influence is carried out by counting the paths that a node is involved in. The average BC values as well as the number of nodes with BC=0 and the highest BC value of each RPM are shown in Table SM-3. Figure SM-2 shows a chart with the highest BC values for the most salient DMOZ models.

Table SM-3: Betweenness centrality values of DMOZ models

Model	Average BC	Number of nodes with BC=0	Highest BC
T	22	435,641	529,605
S	44	512,596	963,578
R	25,121	517,718	105,026,375
G_1	2,708,969	233,362	25,192,567,132
G_2	2,538,287	233,061	14,572,372,601
G_3	0	571,148	0
G_4	1,731,073	264,413	58,235,337,597
G_5	0	571,148	0
G_6	852,643	481,642	3,191,455,243
G_7	26,705	369,333	374,575,132

G_8	22,456	413,026	368,145,533
G_9	784,937	393,117	2,749,380,502
G_{10}	1,183,409	382,971	3,678,132,867
G_{11}	947,509	504,098	34,582,303,776
G_{12}	1,185,520	443,436	39,545,756,469
G_{13}	4,823	486,903	356,494,435
G_{14}	1,274,742	369,141	43,544,454,702
G_{15}	1,050,380	285,249	7,910,607,336
G_{18}	1,274,819	274,422	31,917,718,286
G_{19}	1,078,864	411,080	29,338,006,924
G_{20}	1,732,976	258,773	55,680,921,822

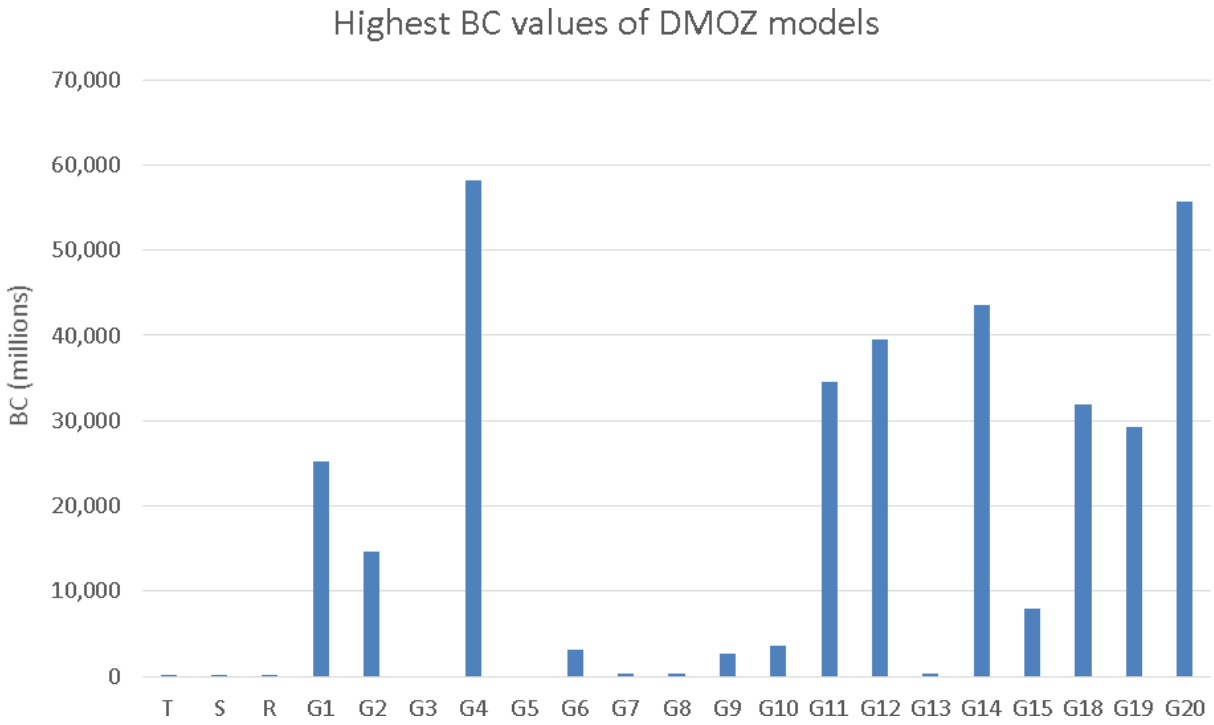

Figure SM-2: Highest BC values in millions for the most important DMOZ RPM's.

The calculation of BC values for some RPM's is computationally very expensive. For instance, it can be observed that model G_4 has a node that participates in more than 58 billion of shortest paths. As seen in Figure SM-3, the analysis of the BC distribution for many RPM's indicates that all of them are heavy-tailed but only the BC values for S are consistent with a power law distribution. Another interesting question about the BC measure in RPM's is the very high magnitude of the difference between the average and highest value.

Besides, it is important to highlight the existence of many alternative paths between some pairs of nodes in a network having the shortest length.

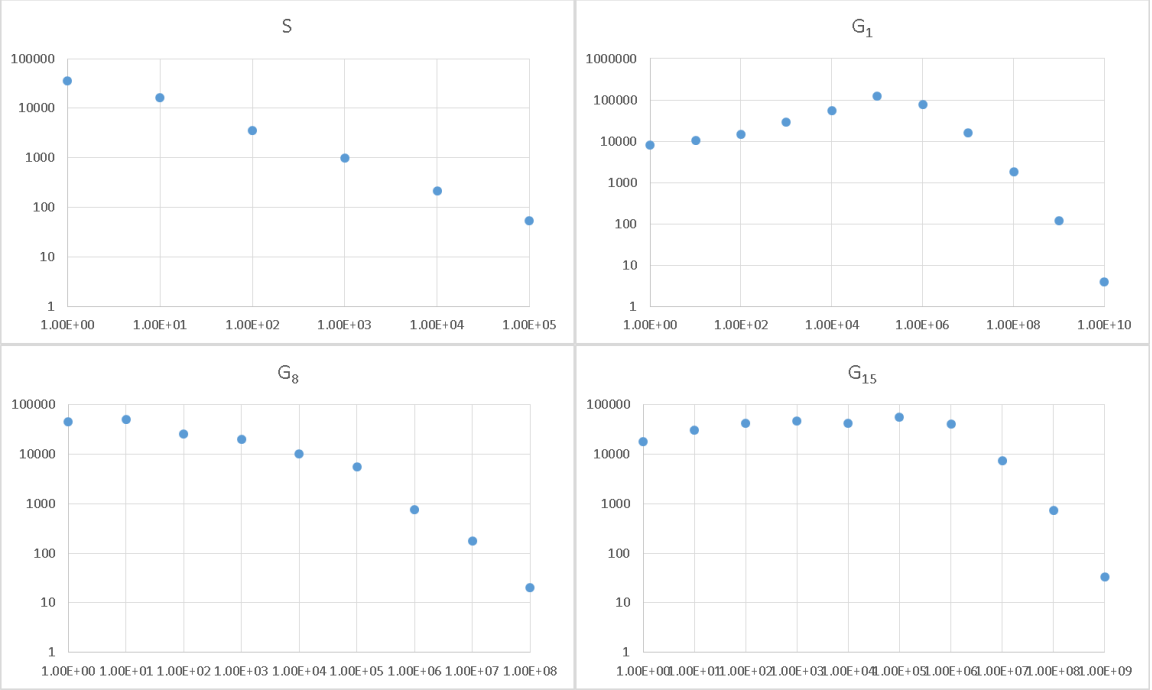

Figure SM-3: BC distribution in log-log scale for some representative DMOZ RPM's.

Alternative Measures of Centrality

This section reports alternative measures computed on the RPM's of DMOZ, with the purpose of assessing their connectivity patterns from another point of view. Table SM-4 shows the mean and max values for the measures of Closeness Centrality, Harmonic Centrality and Lin's Index described in the Background section.

Table SM-4: Average values of Closeness Centrality, Harmonic Centrality and Lin's Index for DMOZ models.

Model	Closeness Centrality		Harmonic Centrality		Lin's Index	
	Avg	Max	Avg	Max	Avg	Max
T	0.042939	1	2.57	3.25	1.74	1.87
S	0.256610	1	1.77	589.92	1.51	520.00
R	0.077657	1	114.28	16179.53	109.03	14441.91
G_1	0.000142	1	25409.22	43242.16	24430.69	39754.73
G_2	0.000002	1	29271.97	47432.96	28145.28	43589.30
G_3	0.152510	1	7.03	14.00	7.03	14.00
G_4	0.000355	1	38280.34	47609.34	36027.50	42373.23
G_5	0.284470	1	5.68	747.00	5.68	747.00

G_6	0.034762	1	7488.22	31691.50	7152.99	29411.45
G_7	0.010165	1	100.50	2244.37	85.97	1816.23
G_8	0.023856	1	121.80	2311.87	99.50	1912.68
G_9	0.151940	1	4429.86	31747.85	4246.06	28761.93
G_{10}	0.088250	1	11790.05	38631.06	11304.51	35755.26
G_{11}	0.000631	1	39473.65	51527.90	37285.34	46391.14
G_{12}	0.000002	1	47103.86	60620.70	44200.87	54468.84
G_{13}	0.026235	1	386.70	5592.87	353.83	4903.17
G_{14}	0.000369	1	49582.84	64048.37	46655.90	57775.42
G_{18}	0.000002	1	53696.74	67034.06	50387.85	60696.77
G_{19}	0.000002	1	61123.58	81575.45	57629.46	71544.55
G_{20}	0.000002	1	41128.94	50922.54	38894.88	45593.31

As denoted in Table SM-4 and explained in (Xamena et al 2017), all the RPM's exhibit a maximum Closeness Centrality value of 1. This is caused by the children of the root node in the main taxonomy, that have a unique co-reachable node, in fact, the root node. The remaining RPM's reflect the presence of, on the one hand, nodes of the taxonomy with the mentioned feature or, on the other hand, nodes that only have one co-reachable node in the S and R components. In the charts of Figure SM-4 and SM-5, very different average values are immediately identified. For the case of Harmonic centrality and Lin's index, many instances have maximum values that are one or two orders of magnitude higher than the mean values. This can be caused by the fact that the corresponding networks have a small number of central nodes, according to those measures. A deeper study over specific central nodes can shed light on the importance of the corresponding DMOZ topics in each RPM.

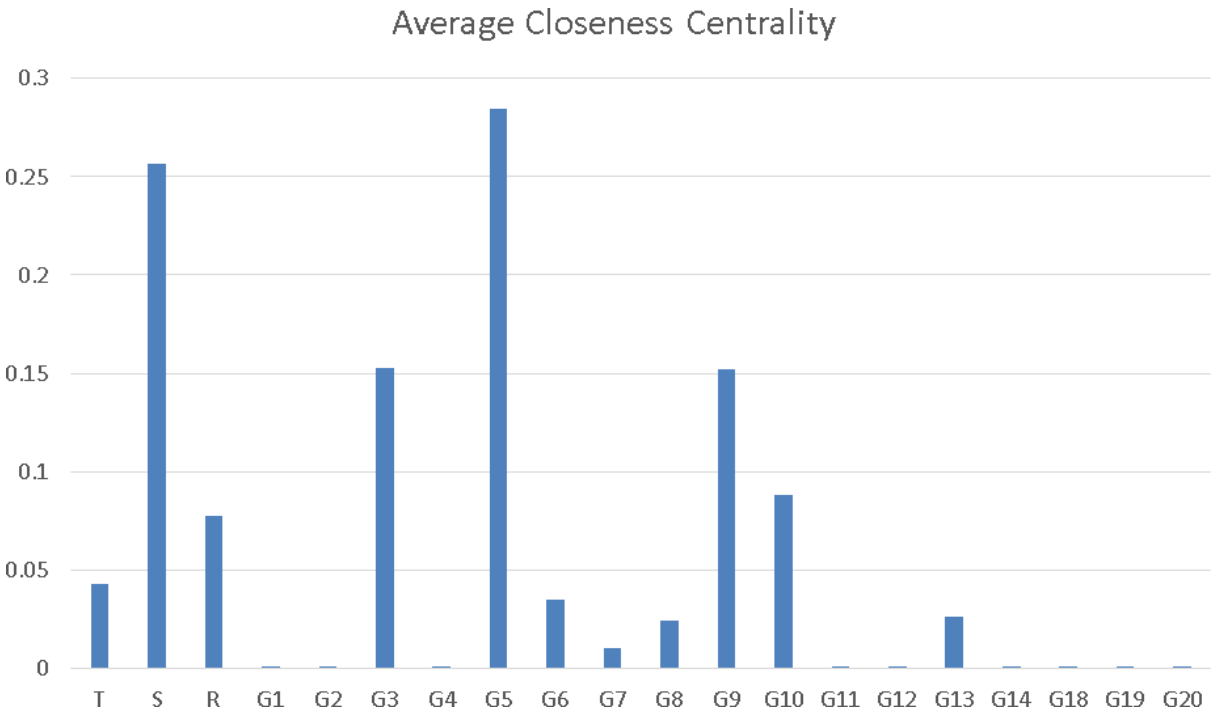

Figure SM-4: Average Closeness Centrality of DMOZ RPM's.

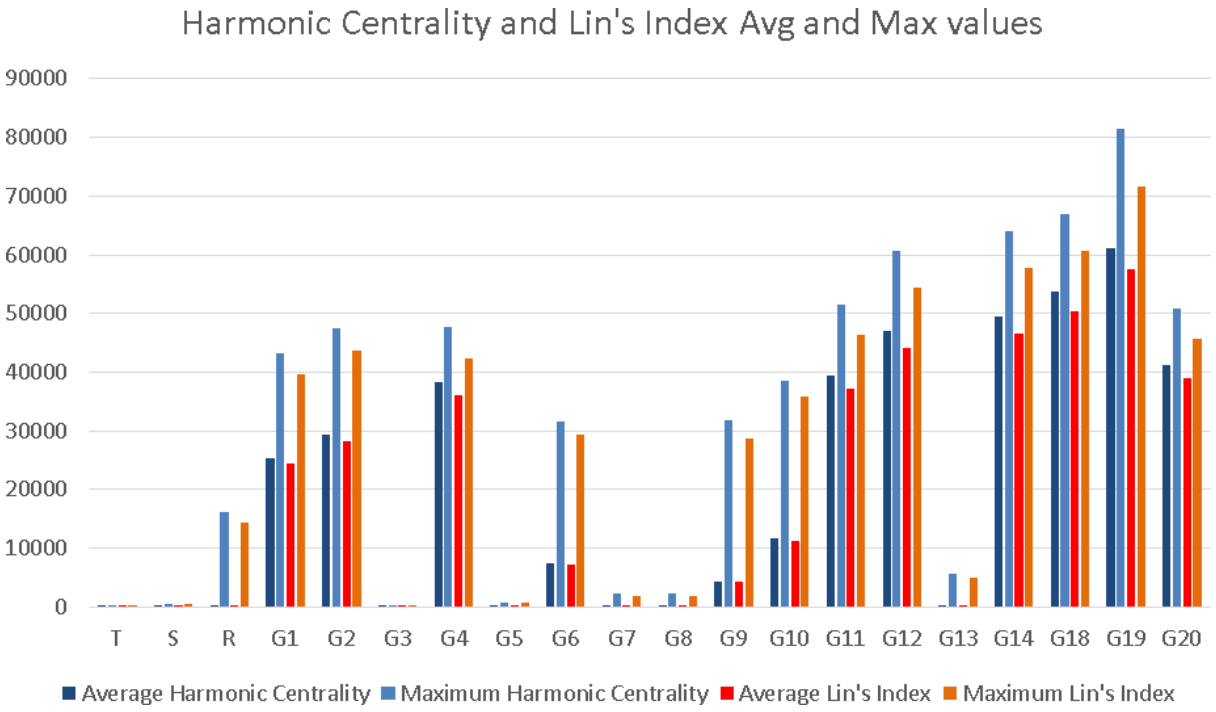

Figure SM-5: Average and maximum values of Harmonic Centrality and Lin's Index of DMOZ RPM's.

References

J. I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, A. Vespignani, *Large scale networks fingerprinting and visualization using the k-core decomposition, in: Advances in neural information processing systems, 2005, pp. 41–50.*

A. Bavelas, *Communication patterns in task-oriented groups., Journal of the acoustical society of America (1950).*

U. Brandes, *A faster algorithm for betweenness centrality*, Journal of mathematical sociology 25 (2) (2001) 163–177.*

L. C. Freeman, *A set of measures of centrality based on betweenness, Sociometry (1977) 35–41.*

N. Lin, *Foundations of social research, McGraw-Hill New York, 1976.*

M. Marchiori, V. Latora, *Harmony in the small-world, Physica A: Statistical Mechanics and its Applications 285 (3) (2000) 539–546.*

M. E. Newman, *Power laws, pareto distributions and zipf's law, Contemporary physics 46 (5) (2005) 323–351.*

S. B. Seidman, *Network structure and minimum degree, Social networks 5 (3) (1983) 269–287.*

D. J. Watts, S. H. Strogatz, *Collective dynamics of "small-world" networks, Nature 393 (6684) (1998) 440–442.*