Supplementary Materials for Navigating concepts in the human mind unravels the latent geometry of its semantic space

Barbara Benigni[1,2*], Monica Dallabona[3], Elena Bravi[3], Stefano Merler[4], Manlio De Domenico[2*]

[1]*Department of Information Engineering and Computer Science, University of Trento, Via Sommarive, 9, 38123 Povo (TN), Italy*

[2]*CoMuNe Lab,* [4]*DPCS , Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo (TN), Italy*

[3]*Department of Mental Health, Division of Psychology, Azienda Provinciale per i Servizi Sanitari, Viale Verona, 38123 Trento, Italy*

Corresponding author: mdedomenico@fbk.eu, bbenigni@fbk.eu

1 Extended Results

Geometry Tables 1 and Tables 2 shows respectively the results of Kolmogorov-Smirnov statistical tests and t-tests on the five local metrics computed using the cosine distance, for each pair of groups, for the three geometries.

		Itwac		Twitter		Wikipedia	
Metrics	Pairs	Test statistic	P-val adj	Test statistic	P-val adj	Test statistic	P-val adj
DOE	DEM-MCI	0.198	0.0526	0.177	1.1002e-01	0.220	2.3099e-02
	DEM-CTR	0.601	4.6765e-07	0.510	4.6160e-05	0.684	3.8300e-09
	MCI-CTR	0.559	2.7697e-06	0.418	1.4290e-03	0.629	6.401564e-08
ρ_w	DEM-MCI	0.186	8.1981e-02	0.254	5.2124e-03	0.219	2.3320e-02
	DEM-CTR	0.657	1.9492e-08	0.609	3.1449e-07	0.791	2.9786e-12
	MCI-CTR	0.647	2.2213e-08	0.501	4.5219e-05	0.720	2.3783e-10
Max_J	DEM-MCI	0.110	0.6293	0.067	0.9859	0.197	5.5572e-02
	DEM-CTR	0.323	0.0532	0.343	0.0295	0.510	4.6160e-05
	MCI-CTR	0.303	0.0617	0.328	0.0304	0.355	1.3224e-02
d	DEM-MCI	0.242	9.047e-03	0.222	2.0896e-02	0.295	6.4264e-04
	DEM-CTR	0.793	1.699e-12	0.760	2.6543e-11	0.837	1.0325e-13
	MCI-CTR	0.748	1.4918e-12	0.686	1.4214e-10	0.773	1.0325e-13
far	DEM-MCI	0.110	0.6320	0.099	0.7552	0.187	7.9703e-02
	DEM-CTR	0.379	0.0060	0.313	0.0712	0.498	5.3973e-05
	MCI-CTR	0.385	0.0047	0.298	0.0712	0.500	3.0045e-05

Table 1: Results of Kolmogorov-Smirnov statistical tests for the three semantic spaces, p-values are adjusted according to Holm–Bonferroni method.

Hierarchy Figure 1 report boxplots within violin plots of the explorative potential distributions, both in terms of visited clusters and in terms of words contained in the visited clusters, for the three categories of subjects in the three geometries. Results of Kolmogorov-Smirnov statistical tests and t-tests for these indicators, for each pair of groups, are shown in tables 3 and 4 respectively.

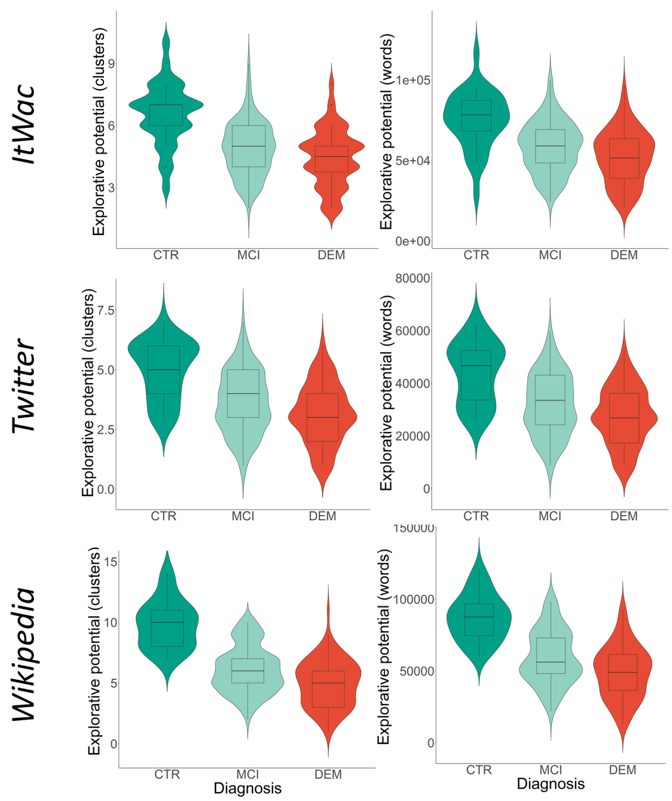

Figure 1: *Explorative potential.* Boxplots within violin plots of the explorative potential distributions, expressed both in terms of visited clusters and of words cointained in the visited clusters for the three categories of subjects in the three semantic spaces. This figure has been generated using the publicly available R software `https://www.r-project.org/`.

		Itwac				Twitter				Wikipedia			
Metrics	Pairs	Test statistic	df	P-val adj	Effect size	Test statistic	df	P-val adj	Effect size	Test statistic	df	P-val adj	Effect size
	DEM-MCI	-1.631	158.613	1.0478e-01	0.240	-1.795	170.281	7.4416e-02	0.264	-2.930	161.536	3.8752e-03	0.432
DOE	DEM-CTR	-6.360	75.738	4.1587e-08	1.088	-6.576	115.991	4.3444e-09	0.933	-9.786	119.985	1.6696e-16	1.320
	MCI-CTR	-5.822	49.928	8.3157e-07	1.210	-5.588	97.287	4.1935e-07	0.861	-8.648	106.536	1.1815e-13	1.282
	DEM-MCI	-2.213	173.545	2.8188e-02	0.326	-2.517	181.439	1.2717e-02	0.370	-2.390	182.351	1.7867e-02	0.351
ρ_w	DEM-CTR	-8.967	71.030	8.0622e-13	1.575	-9.790	95.304	1.3534e-15	1.528	-13.584	59.023	2.4172e-19	2.592
	MCI-CTR	-7.780	55.666	3.6978e-10	1.523	-7.619	87.600	5.6585e-11	1.224	-11.344	63.243	1.3758e-16	2.086
	DEM-MCI	-1.030	172.662	0.3044	0.152	-0.996	174.718	0.3205	0.147	-2.828	173.288	5.2360e-03	0.416
Max_J	DEM-CTR	-4.044	63.191	0.0004	0.747	-3.518	78.970	0.0022	0.592	-6.585	90.702	8.7001e-09	1.048
	MCI-CTR	-3.515	49.862	0.0019	0.731	-2.898	62.437	0.0104	0.536	-4.452	70.443	6.2410e-05	0.781
	DEM-MCI	-2.845	182.467	4.9543e-03	0.418	-2.701	182.213	7.5553e-03	0.397	-3.078	182.098	2.4032e-03	0.453
d	DEM-CTR	-11.520	44.536	1.7978e-14	2.587	-10.584	46.599	1.6573e-13	2.306	-11.865	44.799	6.0547e-15	2.654
	MCI-CTR	-9.850	43.051	2.6911e-12	2.260	-8.953	44.535	3.2412e-11	2.004	-10.087	42.809	1.3955e-12	2.324
	DEM-MCI	-0.464	163.388	0.6430	0.068	-0.725	171.673	0.4696	0.107	-1.932	165.562	5.5037e-02	0.285
far	DEM-CTR	-3.473	109.890	0.0015	0.509	-2.328	118.777	0.0649	0.323	-6.509	87.885	1.3534e-08	1.049
	MCI-CTR	-3.666	79.863	0.0013	0.611	-1.839	106.042	0.1373	0.273	-5.544	61.444	1.3223e-06	1.033

Table 2: Results of t-tests for the three semantic spaces, df stands for degrees of freedom, p-values are adjusted according to Holm–Bonferroni method, the effect size is the value of Cohen’s d.

Network Values of correlation between the steady state distribution ($\vec{\pi}$) of the three groups in the three geometry are shown in tables 5 to 7, while values of mean first passage time matrices (MFPT) correlation are reported in tables 8 to 10. Finally, Frobenius norms of mean first passage time matrices of the three groups, in the three geometries, are reported in table 11.

2 Choice of teleportation parameter in the PageRank algorithm

In 2007, Griffiths, Steyvers and Firl, in their paper "Google and the Mind - Predicting Fluency With PageRank" [1] showed that Page Rank algorithm predicts human response in a fluency task. The parallelism between the google search engine – and more in general the World Wide Web – and the mind lies in the ability to retrieve the information which is relevant to a particular query. The order in which this information is retrieved and thus connected, in the human mind (e.g. concepts), is similar to the way in which Web pages are connected. Thank to this pair-wise association of concepts in human mind is possible to build semantic networks, which have proven to have properties similar to those of the World Wide Web. The most relevant of this property is the "scale-free" degree distribution (Steyvers & Tenenbaum, 2005 [2]). In their work of 2007, Griffiths, Steyvers and Firl [1], with a sort of mimic of the google search engine, aimed to discover which words is most likely to be produced in a fluency task. By comparing Page Rank and other standard predictors computed on a semantic network, they found out that Page Rank outperforms other metrics in predicting the words that people produce during a verbal fluency task. For this reason, they claim that Page Rank of a word could be use in the design or in the model of memory experiments.

Furthermore, taking inspiration from the process of clustering and switching when retrieving concepts from memory, network scientists provided a new kind of random walk over a graph as a Markov process – i.e. the switcher random walk – (Goñi et al., 2010) [3] to generalize the exploration task on a network. In this vein and by following the assumption of a semantic network navigated by a random walk (Abbott, Austerweil& Griffiths, 2015 [4]), we investigated the navigation of concept by means of its Markov chain representation. The rationale behind this representation is given by a parallelism between a random walker walking on a spatial network and a random memory retriever retrieving concepts from a network of concepts, i.e. from a navigation of concepts on top of a network. As it often happens, here the terms “random walk” and “Markov chain” are used interchangeably. For each diagnosis and for the healthy controls we have estimated a Markov chain, i.e. a random walk on a network of concepts, where each states is represented by a cluster of concepts. The Markov chain is represented by a directed graph encoding the semantic network where each state represents a cluster of words and the probability to transit from one state to another is given by a transition matrix. Since we aim at characterizing the exploration of concepts (at this point at the macroscale), we have to evaluate the dynamic of such an exploration on the Markov chain, i.e by considering the steady state distribution and the mean first passage time matrix for each diagnosis. A unique steady-state probability distribution it is guaranteed for any ergodic Markov chain. In order to guarantee the Markov chain to be ergodic – satisfying the conditions of irreducibility and aperiodicity – we modify the transition matrix by adding a damping effect given by the Page Rank algorithm. In formulas, for each category (DEM, MCI, healthy controls) we compute:

$$\widehat{T} = \alpha\widehat{M} + (1-\alpha)\frac{1}{S} \tag{1}$$

Where $\widehat{T}$ represents the new (modified) transition matrix, $\widehat{M}$ is the transition matrix estimated according to the frequency of words pronounced by the subjects belonging to a specific category, α is the damping effect and S is the total number of states of the Markov chain. Moreover, by adding the damping effect we intend to model the navigation of concepts considering two main component that govern the exploration dynamic: a) a word frequency-based component $\widehat{M}$ and b) a random walk uniformly distributed component $(1-\alpha)\frac{1}{S}$. In this way, the second component acts as a sort of noise introduced when modelling the exploration of concepts also to avoid possible overfitting of the model to our data. Relying on the parallelism between google search engine and memory retrieval tasks, among all possible values between 0 and 1 the damping factor is usually set at 0.85 (Brin and Page, 1998 [5], and Mihalcea, Tarau, Figa, 2004 [6], in the field of semantic networks) and this is also the value we arbitrary choose to modify the transition matrix, for each of the three categories.

Curiously, in 1995, three years before Page Rank paper was published by Brin and Page, two cognitive and linguistic scientist, Bradley Love and Steven Sloman [7], proposed an algorithm of centrality equivalent to the Page Rank to measure the features centrality of a given node on a graph for human concepts (this is pointed out also by Griffiths, Steyvers and Firl, 2007 [1]). This last curiosity strengthens the close relationship between the information retrieval processes in the mind and in the World Wide Web as well as it points out that, not surprising, these two different fields of study have proposed equivalent strategies to meet the same purposes, independently.

		Itwac		Twitter		Wikipiedia	
Metrics	Pairs	Test statistic	P-val adj	Test statistic	P-val adj	Test statistic	P-val adj
	DEM-MCI	0.116	0.5617	0.147	0.2736	0.207	3.7673e-02
$clusters$	DEM-CTR	0.560	4.0993e-06	0.514	0.000037	0.824	2.747802e-13
	MCI-CTR	0.444	5.2066e-04	0.402	0.0026	0.696	1.164886e-09
	DEM-MCI	0.169	0.1424	0.193	6.4397e-02	0.242	0.009
$words$	DEM-CTR	0.537	0.00001	0.525	2.3420e-05	0.770	1.3829e-11
	MCI-CTR	0.420	0.0013	0.425	1.1165e-03	0.581	9.1098e-07

Table 3: Results of Kolmogorov-Smirnov statistical tests for the three semantic space, p-values are adjusted according to Holm–Bonferroni method.

		Itwac				Twitter				Wikipiedia			
Metrics	Pairs	Test statistic	df	P-val adj	Effect size	Test statistic	df	P-val adj	Effect size	Test statistic	df	P-val adj	Effect size
	DEM-MCI	-2.255	182.178	2.533251e-02	0.332	-2.590	182.831	1.0359e-02	0.381	-4.023	182.954	8.4070e-05	0.592
$clusters$	DEM-CTR	-6.489	48.141	1.325017e-07	1.385	-7.249	50.613	6.8850e-09	1.501	-11.845	48.600	1.853765e-15	2.513
	MCI-CTR	-5.057	45.848	1.459508e-05	1.110	-5.307	52.294	4.6125e-06	1.074	-9.098	48.196	9.8115e-12	1.933
	DEM-MCI	-2.551	182.138	1.1560e-02	0.375	-2.602	182.394	1.0040e-02	0.382	-3.801	182.997	1.9594e-04	0.559
$words$	DEM-CTR	-5.648	50.317	2.257816e-06	1.174	-6.923	47.106	3.1682e-08	1.498	-10.923	56.615	4.3996e-15	2.128
	MCI-CTR	-3.943	47.709	5.2426e-04	0.843	-5.041	49.658	1.3079e-05	1.051	-7.946	56.743	1.7432e-10	1.540

Table 4: Results of t-tests for the three semantic spaces, df stands for degrees of freedom, p-values are adjusted according to Holm–Bonferroni method, the effect size is the value of Cohen's d.

	$DEM-MCI$	$DEM-CTR$	$MCI-CTR$
$Pearson$	0.99	0.94	0.95
$Spearman$	0.88	0.70	0.71
$covariance$	0.01	0.0086	0.0089
$norm$	0.04	0.13	0.12

Table 5: Correlation values between the steady state distributions in *Itwac* semantic space.

	$DEM-MCI$	$DEM-CTR$	$MCI-CTR$
$Pearson$	0.97	0.96	0.99
$Spearman$	0.84	0.57	0.57
$covariance$	0.02	0.01	0.01
$norm$	0.13	0.14	0.05

Table 6: Correlation values between the steady state distributions in *Twitter* semantic space.

	$DEM-MCI$	$DEM-CTR$	$MCI-CTR$
$Pearson$	0.93	0.85	0.86
$Spearman$	0.64	0.68	0.81
$covariance$	0.002	0.002	0.002
$norm$	0.09	0.16	0.15

Table 7: Correlation values between the steady state distributions in *Wikipedia* semantic space.

	$DEM-MCI$	$DEM-CTR$	$MCI-CTR$
$Pearson$	0.99	0.89	0.91
$Spearman$	0.99	0.94	0.94
$covariance$	371.78	342.85	455.55
$norm$	59.92	111.06	89.34

Table 8: Correlation values between the mean first passage time matrices in *Itwac* semantic space.

	$DEM - MCI$	$DEM - CTR$	$MCI - CTR$
$Pearson$	0.98	0.92	0.88
$Spearman$	0.97	0.90	0.82
$covariance$	156.96	192.56	193.67
$norm$	18.45	61.00	63.00

Table 9: Correlation values between the mean first passage time matrices in *Twitter* semantic space.

	$DEM - MCI$	$DEM - CTR$	$MCI - CTR$
$Pearson$	0.90	0.63	0.63
$Spearman$	0.91	0.77	0.80
$covariance$	2005.85	1639.81	1213.20
$norm$	643.41	921.02	788.58

Table 10: Correlation values between the mean first passage time matrices in *Wikipedia* semantic space.

	CTR	MCI	DEM
$itWac$	1480.51	635.05	389.04
$Twitter$	1085.24	583.08	280.27
$Wikipedia$	3372.137	2054.467	3066.652

Table 11: Frobenius norm of mean first passage time matrices of the three groups and for the three geometries.

References

1. Griffiths, T. L., Steyvers, M. & Firl, A. Google and the mind: Predicting fluency with pagerank. *Psychological Science* **18**, 1069–1076 (2007).

2. Steyvers, M. & Tenenbaum, J. B. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science* **29**, 41–78 (2005).

3. Goñi, J. *et al.* Switcher-random-walks: A cognitive-inspired mechanism for network exploration. *International Journal of Bifurcation and Chaos* **20**, 913–922 (2010).

4. Abbott, J. T., Austerweil, J. L. & Griffiths, T. L. Random walks on semantic networks can resemble optimal foraging. In *Neural Information Processing Systems Conference; A preliminary version of this work was presented at the aforementined conference.*, vol. 122, 558 (American Psychological Association, 2015).

5. Page, L., Brin, S., Motwani, R. & Winograd, T. The pagerank citation ranking: Bringing order to the web. Tech. Rep., Stanford InfoLab (1999).

6. Mihalcea, R., Tarau, P. & Figa, E. Pagerank on semantic networks, with application to word sense disambiguation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 1126–1132 (2004).

7. Sloman, S. A., Love, B. C. & Ahn, W.-K. Feature centrality and conceptual coherence. *Cognitive Science* **22**, 189–228 (1998).