Supplementary Material for:

The multistep greedy algorithm identifies community structure in real-world and computer-generated networks

Philipp Schuetz and Amedeo Caflisch[1]

[1]*Department of Biochemistry, University of Zurich,*

Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

Fax: +41 44 635 68 62, Email: caflisch@bioc.uzh.ch

(Dated: July 4, 2008)

Contents

I. PROPERTIES OF COMPUTER-GENERATED NETWORKS

The diversity among the computer-generated networks is illustrated in Table I. To complement the main text (where only $l = 0.251\sqrt{L}$ is tested with L the number of edges), the modularity $Q_{\rm pred}$ results are displayed for the MSG-VM application with l as in Eq. (2) (six values are tested, i.e. $l_\alpha = \alpha\sqrt{L},\ \alpha = 0.25, 0.5, 0.75, 1$ and the two adjacent integers to l_α yieldings the highest Q value). The value $Q_{\rm pred}$ is smaller than the expectation value $\langle Q_{\rm rand}^{l<1.5\sqrt{L}}\rangle$ only for 32 (of 300), 19 (of 200), and 15 (of 300) networks of type *SED*, *SLD*, and *LLD*, respectively.

Network					MSG-VM with optimal l				MSG-VM with l from Eq. (2)			MSG-VM with random l	
Type	Realization Index	Vertices	Edges (L)	$\bar{k} \pm \sigma$	$l_{opt}/\sqrt{L}$	Q_{opt}	C_{opt}	$\bar{n} \pm \sigma$	Q_{pred}	C_{pred}	$\bar{n} \pm \sigma$	$\langle Q_{\text{rand}} \rangle$	$\langle Q_{\text{rand}}^{l<1.5\sqrt{L}} \rangle$
SED	1	264	284	2.2 ± 4.0	1.19	0.789	22	12 ± 8.5	0.789	22	12.0 ± 9.0	0.789	0.789
SED	2	467	486	2.1 ± 1.1	0.59	0.878	20	23.3 ± 7.6	0.878	20	23.3 ± 7.6	0.878	0.878
SED	3	346	793	4.6 ± 2.4	1.21	0.494	12	28.8 ± 8.4	0.488	11	31.4 ± 5.5	0.484	*0.490*
SED	4	322	817	5.1 ± 14.8	0.94	0.365	10	32.2 ± 18.3	0.364	9	35.7 ± 15.4	0.358	0.363
SED	5	550	942	3.4 ± 14.7	0.68	0.502	16	34.3 ± 29.5	0.502	16	34.3 ± 29.5	0.497	0.501
SED	6	301	1299	8.6 ± 11.5	0.31	0.290	8	37.6 ± 16.4	0.290	8	37.6 ± 16.4	0.284	0.285
SED	7	774	1579	4.1 ± 3.2	0.20	0.531	17	45.5 ± 8.9	0.527	17	45.5 ± 11.3	0.521	0.526
SED	8	636	1699	5.3 ± 17.4	0.19	0.388	12	53 ± 27.3	0.387	14	45.4 ± 23.7	0.379	0.385
SED	9	726	2208	6.1 ± 12.4	0.53	0.377	13	55.8 ± 27.3	0.377	13	55.8 ± 27.3	0.369	0.375
SED	10	513	2716	10.6 ± 23.7	1.30	0.232	8	64.1 ± 23.9	0.230	9	57.0 ± 17.1	0.228	0.230
SED	11	902	3191	7.1 ± 17.0	0.27	0.348	10	90.2 ± 38.9	0.348	10	90.2 ± 38.9	0.338	0.344
SED	12	657	3601	11.0 ± 11.9	0.32	0.270	9	73 ± 33.1	0.266	8	82.1 ± 33.8	0.261	0.266
SED	13	846	3984	9.4 ± 27.4	0.24	0.261	9	94 ± 40.6	0.261	9	94.0 ± 40.6	0.252	0.257
SED	14	743	4914	13.2 ± 10.9	0.66	0.249	10	74.3 ± 18.5	0.248	9	82.5 ± 25.4	0.241	0.245
SED	15	513	2716	10.6 ± 23.7	1.30	0.232	8	64.1 ± 23.9	0.230	9	57.0 ± 17.1	0.228	0.230
SLD	1	289	940	6.5 ± 7.2	0.23	0.374	9	32.1 ± 13.8	0.374	9	32.1 ± 13.8	0.364	0.368
SLD	2	995	1242	2.5 ± 3.0	0.14	0.768	26	38.2 ± 14.3	0.767	27	36.8 ± 13.1	0.764	0.766
SLD	3	1640	2465	3.0 ± 17.3	0.50	0.613	31	52.9 ± 49.7	0.613	31	52.9 ± 49.7	0.610	0.612
SLD	4	2211	3554	3.2 ± 41.7	0.37	0.445	25	88.4 ± 193	0.445	23	96.1 ± 200	0.443	0.444
SLD	5	878	3841	8.7 ± 5.0	0.26	0.324	9	97.5 ± 35.4	0.324	9	97.5 ± 35.4	0.314	0.320
SLD	6	1540	5226	6.8 ± 44.0	0.72	0.261	12	128.3 ± 88.4	0.260	11	140 ± 113.9	0.257	0.258
SLD	7	1117	5270	9.4 ± 23.0	0.34	0.285	9	124.1 ± 53.3	0.284	10	111.7 ± 69.8	0.276	0.282
SLD	8	485	5348	22.1 ± 31.2	0.23	0.152	5	97 ± 9.4	0.151	6	80.8 ± 26.4	0.145	0.148
SLD	9	1591	5432	6.8 ± 4.9	0.05	0.377	12	132.5 ± 67.9	0.375	14	113.6 ± 26.1	0.365	0.373
SLD	10	2242	7993	7.1 ± 47.3	0.07	0.284	13	172.4 ± 117	0.280	16	140.1 ± 118.8	0.279	0.280
SLD	11	904	11668	25.8 ± 25.0	0.08	0.166	6	150.6 ± 64.4	0.163	6	150.6 ± 34.3	0.156	0.161
SLD	12	721	11865	32.9 ± 33.0	0.39	0.136	7	103 ± 20.2	0.135	6	120.1 ± 52.8	0.129	0.133
SLD	13	1689	15722	18.6 ± 47.5	0.43	0.178	7	241.2 ± 60.1	0.176	6	281.5 ± 55.6	0.168	0.175
SLD	14	2992	25525	17.1 ± 50.4	0.08	0.196	8	374 ± 244.7	0.195	8	374 ± 220.7	0.185	0.193
SLD	15	3393	26257	15.5 ± 8.8	0.20	0.233	10	339.3 ± 246.2	0.231	10	339.3 ± 203.4	0.221	0.229
LLD	1	544	2084	7.7 ± 27.5	0.39	0.236	8	68 ± 29.5	0.235	8	68 ± 28.4	0.232	0.235
LLD	2	438	5061	23.1 ± 28.0	0.84	0.151	7	62.5 ± 11.1	0.151	6	73 ± 9.1	0.146	0.149
LLD	3	1469	6841	9.3 ± 44.0	0.08	0.232	10	146.9 ± 70.9	0.228	11	133.5 ± 66.5	0.226	0.227
LLD	4	594	6818	23.0 ± 27.6	0.33	0.167	7	84.8 ± 34	0.165	6	99 ± 13.7	0.156	0.161
LLD	5	3869	7931	4.1 ± 61.6	0.94	0.367	21	184.2 ± 305.3	0.367	21	184.2 ± 304.8	0.366	0.366
LLD	6	590	8086	27.4 ± 32.1	0.26	0.144	5	118 ± 9.2	0.144	5	118 ± 9.2	0.136	0.140
LLD	7	3554	8609	4.8 ± 58.6	3.21	0.345	22	161.5 ± 205.7	0.344	21	169.2 ± 222.2	0.342	0.343
LLD	8	2281	12298	10.8 ± 58.0	0.04	0.210	10	228.1 ± 97.7	0.207	13	175.4 ± 86.9	0.203	0.206
LLD	9	4193	19992	9.5 ± 66.5	0.11	0.249	12	349.4 ± 212.3	0.248	13	322.5 ± 200.6	0.235	0.244
LLD	10	1002	26622	53.1 ± 35.5	0.28	0.113	6	167 ± 56.1	0.110	6	167 ± 22.6	0.107	0.110
LLD	11	1133	36783	64.9 ± 41.2	0.34	0.101	6	188.8 ± 40.8	0.099	5	226.6 ± 40.5	0.096	0.098
LLD	12	3485	42297	24.3 ± 53.1	0.32	0.166	8	435.6 ± 188	0.164	7	497.8 ± 262.8	0.155	0.163
LLD	13	2335	54802	46.9 ± 55.5	0.34	0.116	7	333.5 ± 67.5	0.116	7	333.5 ± 53.4	0.108	0.114
LLD	14	3010	75776	50.3 ± 41.7	0.16	0.120	6	501.6 ± 174	0.119	7	430 ± 107.8	0.113	0.118
LLD	15	4165	127797	61.4 ± 50.6	0.26	0.108	5	833 ± 113.6	0.107	6	694.1 ± 323.2	0.101	0.106

TABLE I: Heterogeneity of computer-generated networks and comparison of MSG-VM results using l as in Eq. (2) of main text or picked at random. For each of the three network types, 15 realizations are shown ranked by L. The degree heterogeneity is evident in the average and standard deviation of the degree (column "$\bar{k} \pm \sigma$"). The column "Q_{opt}" lists the maximal value of modularity obtained by running MSG-VM for all values of l smaller than $\min\{5000, L\}$ (L the number of edges). The column "Q_{pred}" lists the MSG-VM modularity obtained using Eq. (2) of the main text to determine the step width. The columns "C_{opt}" and "C_{pred}" list the number of communities in the partitions with modularity Q_{opt} and Q_{pred}, respectively. The average and standard deviation of the number of vertices per community are listed in the columns "$\bar{n} \pm \sigma$". The columns "$\langle Q_{\text{rand}} \rangle$" and "$\langle Q_{\text{rand}}^{l<1.5\sqrt{L}} \rangle$" show the expectation value for the MSG-VM modularity when six values of l are picked randomly from a uniform distribution in the range $1 \leq l \leq \min\{5000, L\}$ and $1 \leq l \leq 1.5\sqrt{L}$, respectively. The expectation value is estimated by averaging, over 1000 samples, the highest modularity obtained using six values of l (confer section VII for details). A total of six values of l are picked randomly because six values were used to determine Q_{pred}: the four values of l calculated by Eq. (2) of the main text and the two integers adjacent to the best of these four. There is only one value of $\langle Q_{\text{rand}} \rangle$ and $\langle Q_{\text{rand}}^{l<1.5\sqrt{L}} \rangle$ higher than the corresponding Q_{pred} (in italics).

Label	#	Deg.	$\frac{l_{\text{opt}}}{\sqrt{L}} > 1.5$	VM-effect	MSG-effect
GN_1	100	61	13 %	13	0
GN_2	100	35	15 %	12	3
GN_3	100	71	18 %	18	0
SED	300	54	7.3 %	19	3
SLD	200	23	5.5 %	9	2
LLD	300	10	0.7 %	0	2

TABLE II: Statistical properties of l_{opt} (smallest value of step width that yields the highest MSG-VM modularity) for the computer-generated networks. The column *Deg.* lists the number of networks for which multiple values of l yield $Q_{\text{MSG-VM}}\,(l_{\text{opt}})$. The fraction of examples with $l_{\text{opt}} > 1.5\sqrt{L}$ (L total edge weight) is small. These networks are classified according to the *VM-labels* (confer Sec. III A for details): "VM-effect" and "MSG-effect".

II. SOURCES OF DEGENERACY

For multiple computer-generated networks more than one value of step width yield the highest MSG-VM modularity (Table II). In contrast, all real-world networks with three exceptions (the jazz, the metabolic *E. coli*, and the Zachary karate club network) have a unique optimal value l_{opt} of the step width. For the Girvan-Newman networks $GN_{1,2,3}$ the number of l_{opt} values displays a phase-transition like behavior (Fig. 1) upon variation of z_{out} (average number of edges connecting a vertex in one of the four imposed communities to members of another module). The transition between the two "phases" (low z_{out} value with many l_{opt} values and high z_{out} value with few l_{opt}) occurs for similar values of $\frac{z_{\text{out}}}{\text{max. degree}}$ whereas the fraction becomes larger with increasing network size. In networks with small z_{out} value many vertices and their neighborhoods are similar in contrast to graphs with high z_{out} value. Therefore, this symmetry (almost identity of vertices) is assumed to be the source of degeneracy.

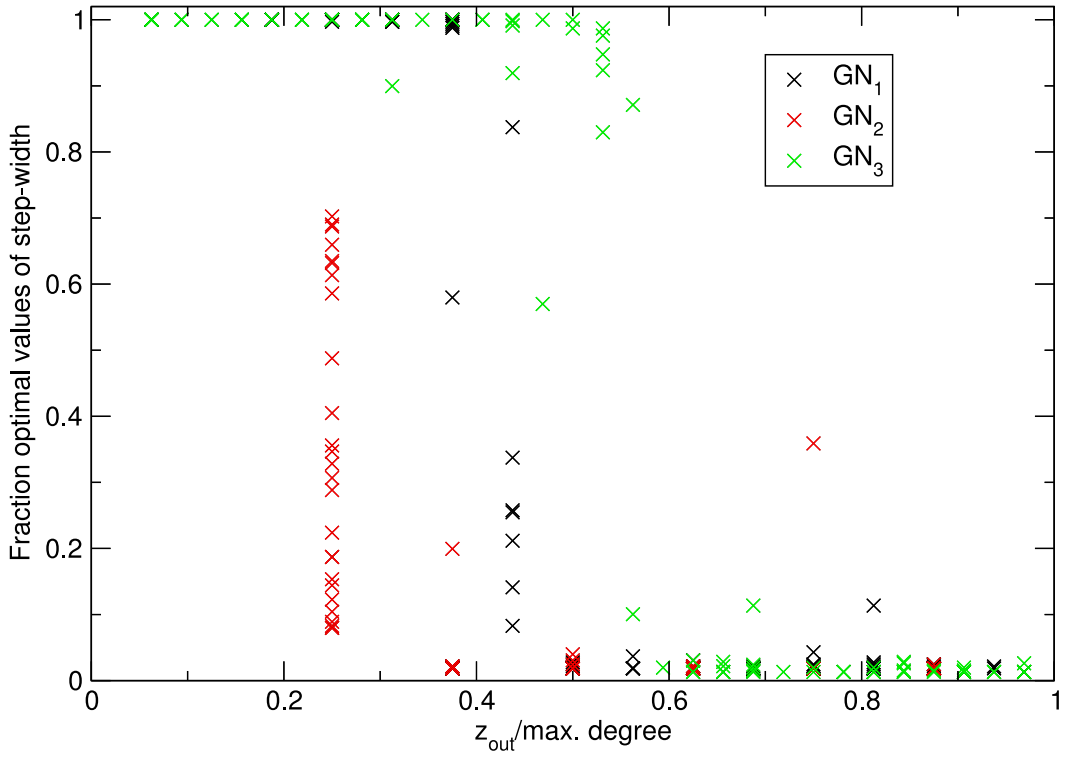

FIG. 1: Influence of z_{out} (average number of inter-community edges per vertex) on the number of distinct values of step width yielding the highest MSG-VM modularity in the Girvan-Newman networks $GN_{1,2,3}$.

III. RESTRICTABILITY OF SEARCH RANGE

A. Quantification of VM-Contribution

The best MSG-VM solution can emerge from two disparate scenarios: First, an excellent MSG solution is insignificantly finetuned by the VM procedure (*MSG-effect*). Second, the VM bears the lion's share and optimizes a poor MSG solution (*VM-effect*). To discern these two cases, the following criteria are defined (l_{opt} is smallest value of the step width that maximizes the MSG-VM modularity, $Q_{\text{MSG}}(l)$ the modularity after the application of the MSG algorithm with l as step width):

"MSG-effect" The modularity obtained by the MSG algorithm with l_{opt} as step width is at least 90 % of the maximal MSG modularity (among all other tested values of step width):

$$Q_{\text{MSG}}(l_{\text{opt}}) > 0.9 \max_{l}(Q_{\text{MSG}}(l))$$

and among the best 50 % of all Q_{MSG}-values:

$$Q_{\mathrm{MSG}}(l_{\mathrm{opt}}) > \frac{\min_l \left(Q_{\mathrm{MSG}}(l)\right) + \max_l \left(Q_{\mathrm{MSG}}(l)\right)}{2}.$$

The second condition assures that no fallacious MSG-dominance is identified when the MSG modularity values fluctuate less than 20 % upon variation of the step width.

"VM-effect" All other cases.

B. Optimal parameter smaller than $1.5\sqrt{L}$

If the MSG algorithm dominates the optimization, the arguments in section III.A of the main text imply that the optimal value of the step width l_{opt} (smallest integer that yields the highest MSG-VM modularity) should be smaller than $\beta\sqrt{L}$ (L total edge weight and β a prefactor on the order of 1). For the VM-driven examples no such concentration can be expected. To test this hypothesis, the computer-generated networks are splitted according to the criteria in section III A. For both groups, a histogram of $r = \frac{l_{\mathrm{opt}}}{\sqrt{L}}$ is calculated (not shown). In agreement with the hypothesis the MSG distribution drops significantly between $r = 1$ and $r = 2$ independent of the network type. Choosing $r_{\mathrm{th}} = 1.5$ as threshold, 92.6 % of the computer-generated networks have $l_{\mathrm{opt}} < r_{\mathrm{th}}\sqrt{L}$. Among these 1019 networks only 41 are VM-driven. 71 of 81 networks with $l_{\mathrm{opt}} > r_{\mathrm{th}}\sqrt{L}$ are VM-driven (Table II). The cutoff value $r_{\mathrm{th}} = 1.5$ demarcates the boundary between MSG- and VM-driven networks. Furthermore, the optimal value of the step width is expected to be an integer smaller than $1.5\sqrt{L}$.

IV. EFFECTS OF VERTEX LABELING PERMUTATION

Permuting the vertex labels leaves the topology invariant, but changes the order in which the MSG-VM algorithm parses the vertices. Thus, the influence of non-topological contributions on the optimization can be studied. Furthermore, by averaging over the profiles of the same network with different vertex labellings the intrinsic accuracy limit of a prediction procedure based on topological properties can be determined. Here, 100 variants of the smallest 10 real-world networks with different vertex labellings are created. On each copy the MSG-VM algorithm is applied for all possible values of step width (i.e. all integers

				VM-Label	
Network	# l_{opt}	>	Δ[%]	VM	MSG
Zachary	78	0	0.00	0	100
Metabolic *E. coli*	49	24	0.12	3	97
College Football	400 [a]	12	0.18	17	83
Metabolic *C. elegans*	141	50	0.91	25	75
Jazz	700 [b]	0	0.00	79	21 [c]
Email	94	80	0.57	15	85
Yeast (PPI, LCC)	18	4	0.05	0	100
M. Karplus	107	36	0.94	4	96
PPI *S. cerevisiae* (LCC)	56	69	0.46	0	100
PPI *S. cerevisiae*	56	82	0.52	0	100

[a]Omitting the networks with "VM-effect" label only 20 different values are found.
[b]If the networks with "VM-effect" are omitted, only six distinct values are found.
[c]Only 15 networks have exclusively l_{opt} values with "MSG-effect" label.

TABLE III: Effect of permutation of vertex labels: For the smallest 10 real-world networks 100 copies with scrambled vertex labels are created. On each copy the MSG-VM procedure is applied using all integers smaller than the number of edges as step width. The column $\#l_{opt}$ lists the number of distinct values of the step width yielding the highest MSG-VM modularity. The number of copies for which the highest MSG-VM modularity is higher than the one for the unscrambled variant is listed in column ">". The highest improvement is shown in column "Δ". The columns "VM-label" report on the number of copies for which the optimization is "MSG" or "VM" dominated.

smaller than the number of edges). The maximal MSG-VM modularity improves at most by 0.94% in comparison to the unscrambled variant (Table III). This change is marginal and smaller than the modularity change upon variation of the step width. Therefore, an empirical formula using topological properties to predict l_{opt} has an accuracy limit of at least one percent.

Strikingly, whether the VM or the MSG dominates the optimization is conserved under vertex label permutation. The excessive diversity of the optimal values of l for the *college football* and the *jazz* network originates from the VM-driven optimization. By taking the

Label	Mean [%]	$\alpha_{\max}$	Median [%]	α_{median}
Zachary Karate Club	100.00	0.57	100.00	0.57
Metabolic *E. coli*	99.90	0.047	99.92	0.47
College Football	98.85	0.080	99.82	0.040
Metabolic *C. elegans*	98.91	0.71	99.00	0.71
Jazz	99.96	0.57	99.97	0.52
Email	99.51	0.65	99.56	0.65
Yeast (PPI, LCC)	99.37	0.27	99.43	0.27
M. Karplus	99.10	0.72	99.17	0.72
PPI *S. cerevisiae* (LCC)	99.55	0.48	99.60	0.48
PPI *S. cerevisiae*	99.49	0.49	99.56	0.49

TABLE IV: Effect of permuted vertex labels on location and value of the peak in the MSG-VM modularity curve. The average/median of the MSG-VM modularity profiles $\bar{Q}_{\text{MSG-VM}}$ is calculated over 100 copies of the smallest ten real-world networks networks with permuted vertex labels. For the mean and median curve the peaks are located at the relative parameter $\alpha_{\max}$ and α_{median}, respectively.

mean or the average over the $\bar{Q}_{\text{MSG-VM}}$ (α) profiles (defined in Sec. III.A.1. of the main text) of the 100 variants with scrambled vertex labels the non-topological contributions are smoothed out. Remarkably, the resulting average/median profiles peak for almost all networks (two exceptions with $l_{\text{opt}} = 1$ and the email network) at values of α in close vicinity of multiples of 0.25 (Table IV).

V. STABILITY OF SCALING FACTOR

The MSG-VM algorithm performs best (on average over all computer-generated networks), if $\lfloor \alpha\sqrt{L} \rfloor$ with $\alpha = 0.251$ is chosen as value of the step width. To assess the effect of another selection of the network set, a leave-23-out test is performed. For this test 10000 samples of 980 out of 1003 computer-generated networks (the 97 networks for which the MSG-VM modularity is independent of the chosen step width are excluded) and for each sample $\bar{Q}_{\text{MSG-VM}}(\alpha)$ is calculated. In Fig. 2 the average and standard deviation

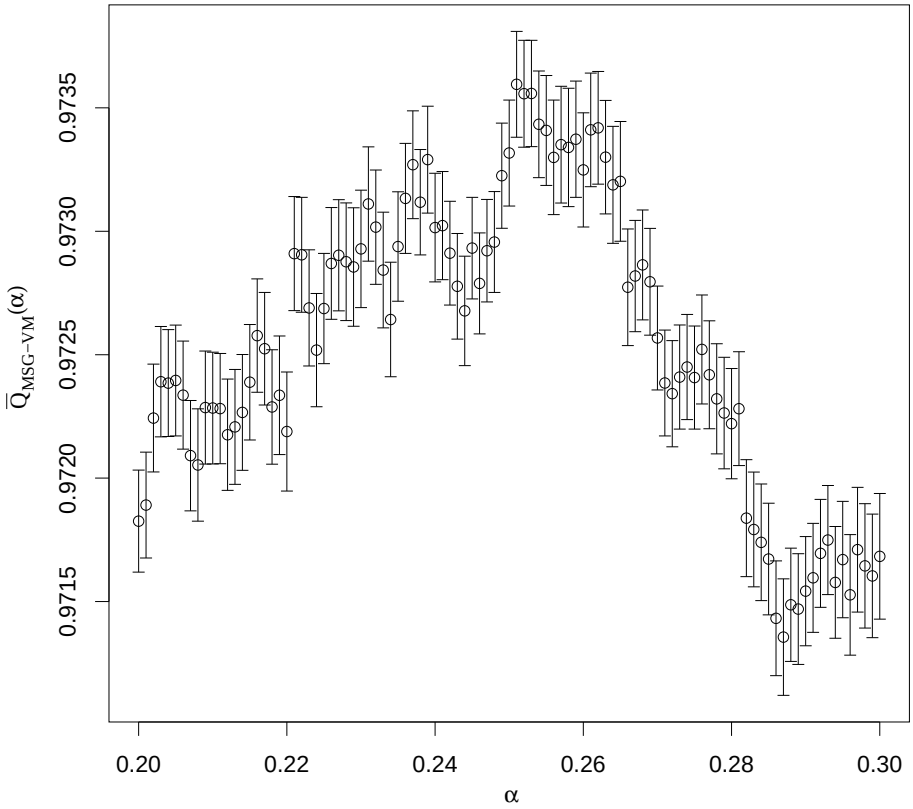

FIG. 2: Leave-23-out procedure on $\bar{Q}_{\text{MSG-VM}}(\alpha)$ curve (definition in Sec. III.A.1 of main text): 10000 samples of 980 out of 1003 computer-generated networks (97 network are excluded as they display no dependence on different values of the step width) are taken and the statistics over the corresponding $\bar{Q}_{\text{MSG-VM}}(\alpha)$ curves is displayed. The peak is at $\alpha = 0.251$. Furthermore, for $\alpha = 0.252, 0.253$ the one-σ-ranges overlap considerably.

of all $\bar{Q}_{\text{MSG-VM}}$ curves are shown. Remarkably, the average $\bar{Q}_{\text{MSG-VM}}$ (average over 10000 samples) peaks at $\alpha = 0.251$, too. At $\alpha = 0.252, 0.253$ the average $\bar{Q}_{\text{MSG-VM}}(\alpha)$ values are within the standard deviation of the values at $\alpha = 0.251$. Reminding that $\alpha = 0.251, 0.252, 0.253$ predict basically the same value of the step width for networks with less that 10^6 edges, this variance in α is negligible.

VI. DETAILS OF Q_{pred} CALCULATION

The modularity Q_{pred} in Table I of the main text is calculated as

$$Q_{\text{pred}} = Q_{\text{MSG-VM}}\left(\left\lfloor \alpha\sqrt{L} \right\rfloor + l_\alpha\right)$$

Network	α	Δl_α
Zachary Karate Club	0.25, 0.5, 0.75	+1
Metabolic *E. coli*	0.25	0
College Football	0.75, 1	−1
Metabolic *C. elegans*	0.75	−1
Jazz	0.75, 1	0
Email	0.75	+1
Yeast (PPI, LCC)	1	0
M. Karplus	0.75	0
PPI *S. cerevisiae* (LCC)	0.5	+1
PPI *S. cerevisiae*	0.5	−1
Internet	0.5	−1
PGP-key signing	0.25	+1
Word Association (LCC)	1	−1
Word Association	0.75	+1
Collaboration	0.5	−1
WWW	0.75	0
Actor	0.75	+1

TABLE V: Parameters for $Q_{\text{pred}} = Q_{\text{MSG-VM}}\left(\left\lfloor \alpha\sqrt{L}\right\rfloor + l_\alpha\right)$ calculation (L the number of edges) used in Table I of the main text.

with α,l_α as given in Table V in the supplementary material and L the number of edges.

VII. CALCULATION OF $\langle Q_{\text{rand}}\rangle$ AND $\langle Q_{\text{rand}}^{l<1.5\sqrt{L}}\rangle$

The expectation values for a random selection of the step width (Table I of the main text) $\langle Q_{\text{rand}}\rangle$ and $\langle Q_{\text{rand}}^{l<1.5\sqrt{L}}\rangle$ are not calculated accurately. These values are approximated by the averages over 1000 sampling experiments in which the highest modularity is calculated for six values of the step width picked at random. The six values of the step width are selected from all tested values for $\langle Q_{\text{rand}}\rangle$ and from those values of step width smaller than $1.5\sqrt{L}$ (L total edge weight) for $\langle Q_{\text{rand}}^{l<1.5\sqrt{L}}\rangle$.

VIII. CORRELATIONS OF l_{opt} WITH TOPOLOGICAL PROPERTIES

Table VI displays a selection of correlation values of topological properties and powers thereof with l_{opt} (the smallest value of the step width yielding the highest MSG-VM modularity). In this calculation only the 905 computer-generated network with a unique l_{opt} are included to reduce the effect of ambiguities. The powers $0.1, 0.2, \ldots, 2$ are considered. Higher powers are excluded to reduce the danger of overfitting. Irrespective of the power considered, the quantities "vertices" and "edges" correlate much better with l_{opt} than all other properties.

IX. EXHAUSTIVE COMPARISON OF MSG-VM AND GREEDY PARTITION FOR METABOLIC NETWORK OF *E. COLI*

The differing parts between the MSG-VM and the greedy partition of the metabolic *E. coli* network are shown in Fig. 3 and 4 (identically classified vertices and edges to/among them are removed). The vertices of the pathway "Puridine Metabolism" (green shaded area) are put in one community by both algorithms. However, the greedy algorithm merges these vertices with those around vertex "C00049" (Aspartate) whereas the MSG-VM algorithm separates them. As aspartate is not involved in the "Puridine Metabolisme", this merge performed by the greedy algorithm is artificial. The MSG-VM algorithm gathers almost all metabolites (the exceptional vertex "C00084" stands for "Acetaldehyde") belonging to the "Glycerophospholipide Metabolism" (yellow shaded area) in one community. In contrast, the greedy algorithm separates the vertices "C01233" (Glycerophosphoethanolamine) and "C000093" (Glycerol-3-phosphate) in addition. Acetaldehyde is part of multiple pathways and therefore might be assigned to other pathways as well. The metabolite "C01233" belongs to only one pathway and therefore is misplaced by the greedy algorithm. Glycerol-3-phosphate belongs to the "Glycerolipid metabolism" pathway to which the vertex "C00577" is assigned as well. To summarize, the greedy algorithm merges artificially two modules and misassignes three vertices of which one is uniquely assigned. The MSG-VM approach misplaces for these two examples only one vertex.

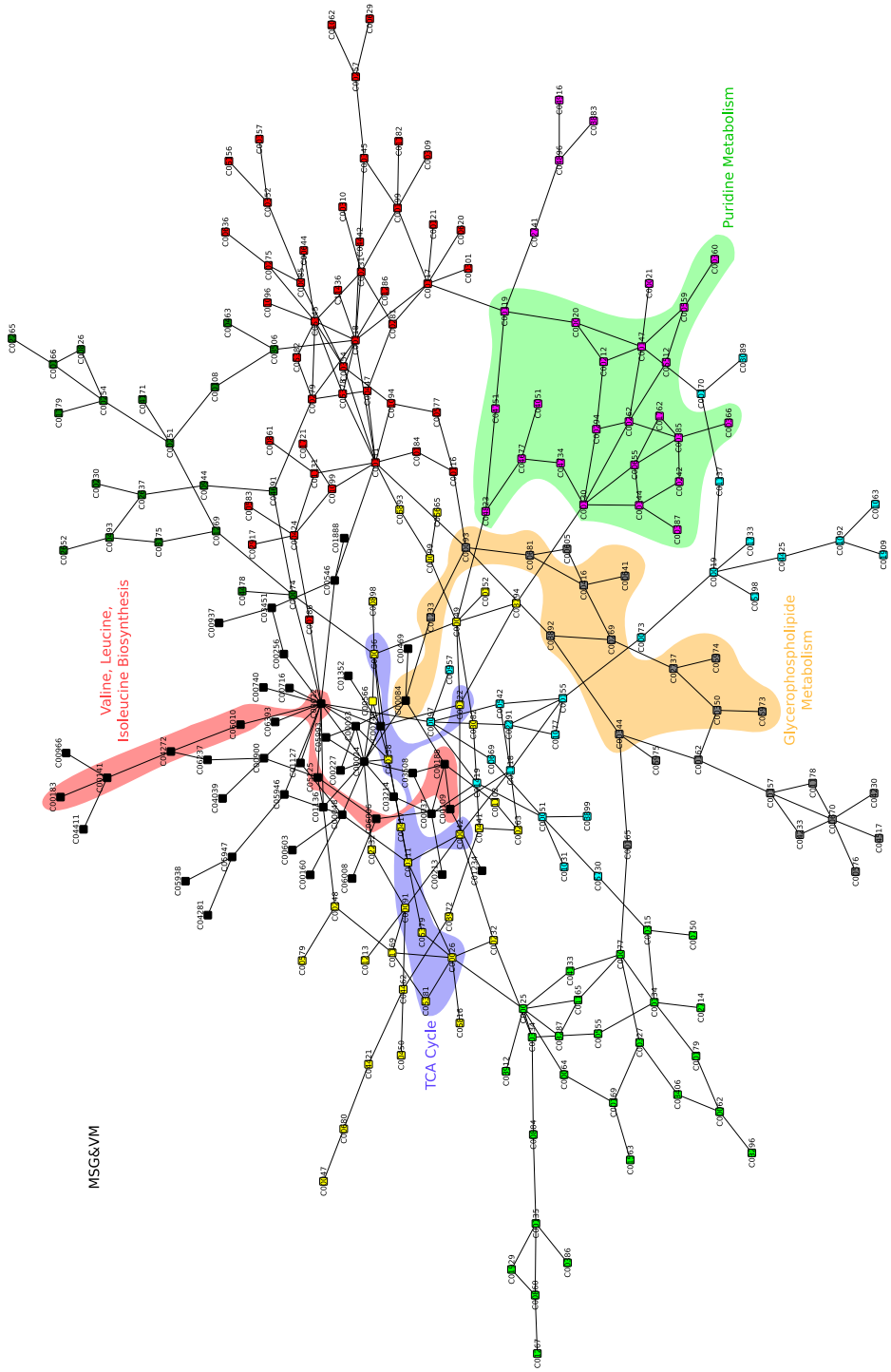

FIG. 3: Partition of the metabolic network of *E. coli* obtained by the MSG-VM algorithm with $l = 6$. The network is reduced on those vertices and edges belonging to communities that differ between the MSG-VM and the greedy solution. The vertex labels are taken from the KEGG database. For instance, detailed information about vertex C00149 can be found at `http://www.genome.jp/dbget-bin/www_bget?compound+C00149`).

	Power			
	0.5	1	1.5	2
Vertices	0.6498	0.6599	0.6482	0.6289
Edges	0.7728	0.7314	0.6669	0.6067
$\langle$Degree$\rangle$	0.6584	0.6596	0.6350	0.6009
Max. Degree	0.4419	0.3589	0.2789	0.2153
σ(Degree)	0.6428	0.6538	0.6384	0.6123
$\langle$CC$\rangle$	-0.0339	-0.0424	-0.0390	-0.0325
σ(CC)	-0.3363	-0.2877	-0.2335	-0.1838
$\langle$CC$^2\rangle$	-0.1021	-0.0583	-0.0362	-0.0266
σ(CC^2)	-0.1944	-0.1876	-0.1647	-0.1366
$\langle$CC$^3\rangle$	-0.1225	-0.0573	-0.0355	-0.0278
σ(CC^3)	-0.1266	-0.1587	-0.1527	-0.1296
$\langle\Delta$CC$\rangle$	-0.3013	-0.2024	-0.1318	-0.0915
$\sigma(\Delta$CC)	-0.4326	-0.3260	-0.2340	-0.1660
$\langle\Delta$Degree$\rangle$	0.2719	0.1318	0.0580	0.0352
$\sigma(\Delta$Degree)	0.2294	0.1117	0.0532	0.0297

TABLE VI: Correlations of topological properties and powers thereof with l_{opt} (value of step width that yields the highest MSG-VM modularity). $\langle x\rangle$ indicates that the average of property x over all vertices is considered. The standard deviation of property y over all vertices is abbreviated by $\sigma(y)$. "CC" stands for clustering coefficient. $\langle\Delta$Degree$\rangle$ and $\langle\Delta$CC$\rangle$ denotes the average over the differences in degree and clustering coefficient of two linked vertices, respectively.

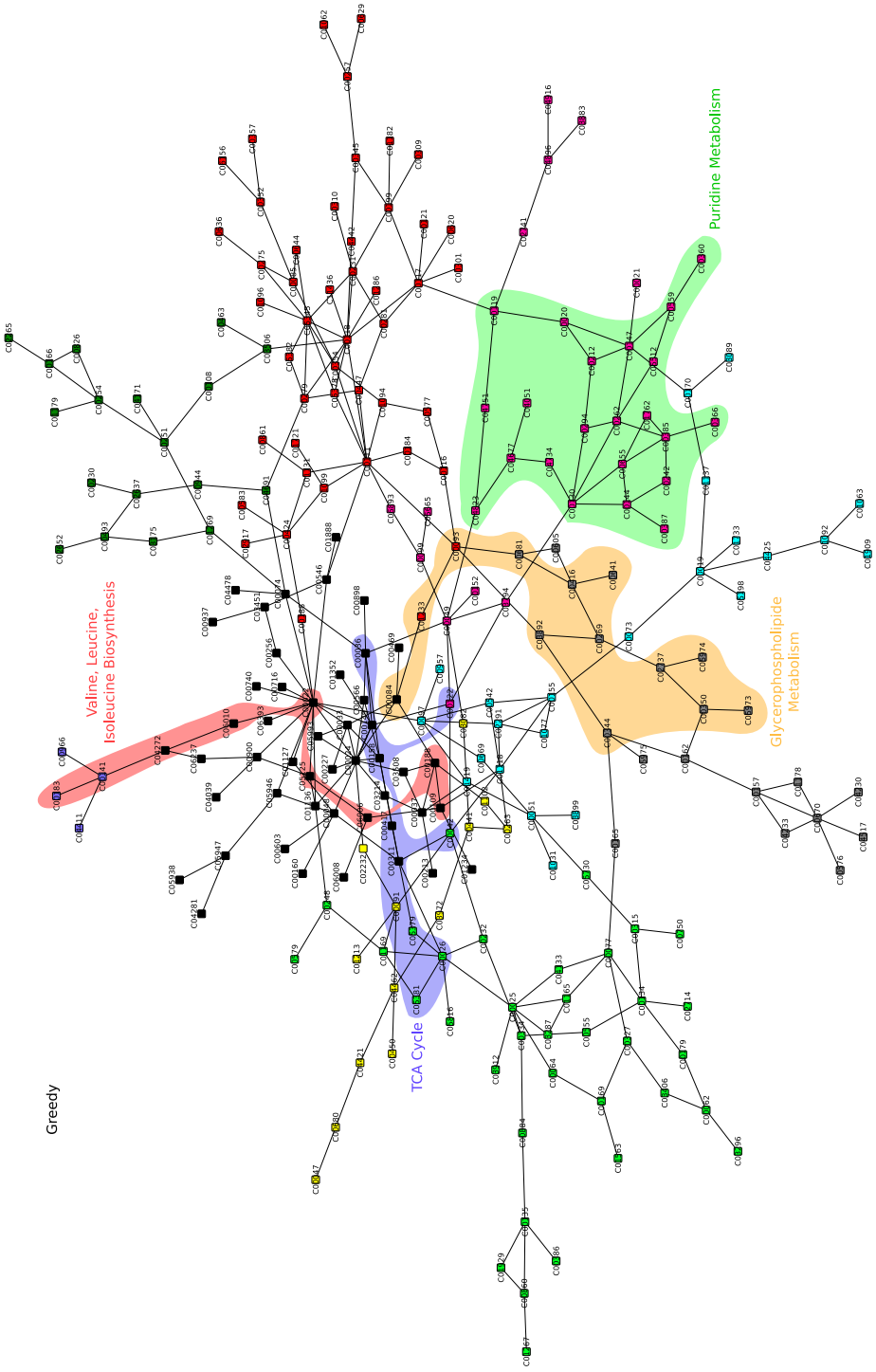

FIG. 4: Clusterization of the metabolic network of *E. coli* obtained by the greedy algorithm. The same excerpt as in Fig. 3 is displayed.