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

Philipp Schuetz* and Amedeo Caflisch†
Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
(Received 18 April 2008; published 20 August 2008)

We have recently introduced a multistep extension of the greedy algorithm for modularity optimization. The extension is based on the idea that merging l pairs of communities ($l>1$) at each iteration prevents premature condensation into few large communities. Here, an empirical formula is presented for the choice of the step width l that generates partitions with (close to) optimal modularity for 17 real-world and 1100 computer-generated networks. Furthermore, an in-depth analysis of the communities of two real-world networks (the metabolic network of the bacterium *E. coli* and the graph of coappearing words in the titles of papers coauthored by Martin Karplus) provides evidence that the partition obtained by the multistep greedy algorithm is superior to the one generated by the original greedy algorithm not only with respect to modularity, but also according to objective criteria. In other words, the multistep extension of the greedy algorithm reduces the danger of getting trapped in local optima of modularity and generates more reasonable partitions.

DOI: 10.1103/PhysRevE.78.026112 PACS number(s): 89.75.Fb, 05.10.−a, 89.75.Kd, 89.75.Hc

I. INTRODUCTION

The coarse-grained organization of many real-world networks manifests itself in a natural divisibility of the vertices into modules (or communities). A community is a set of vertices that are more connected among each other than with vertices of other communities. Community structure has been reported for social networks [1,2], metabolic networks [3–5], and protein-folding networks [6–10]. Several procedures have been developed to partition a network into modules. Often applied are techniques that rely on the optimization of a scoring function called *modularity* [11]. This assessment function compares the fraction of edges within a module with its expectation value in the case of randomly connected vertices with equal degree sequence. The modularity is defined as

$$Q = \sum_{i=1}^{N_C} \left[\frac{I(i)}{L} - \left(\frac{d_i}{2L} \right)^2 \right], \tag{1}$$

with $I(i)$ being the weights of all edges linking vertices of community i, d_i the sum over all vertex degrees in module i, L the total edge weight, and N_C the number of communities. The optimization of modularity has been proven to be a *NP*-hard problem [12]. Thus, heuristic techniques such as extremal optimization [13], simulated annealing [4], and the greedy algorithm [14] have been developed and applied to find partitions with high modularity. Because of the global character of modularity [i.e., in Eq. (1) the connectivity and degree of each community are compared with the edge weight of the whole network], it has been shown that modules smaller than a certain scale cannot be resolved [15]. In other words, small communities are amalgamated with others instead of being detected autonomously. A higher-resolution variant of modularity, called *localized* modularity, addresses the limit on the detectable community size [5].

Recently, we have introduced a multistep extension of the greedy algorithm (MSG) and combined it with a simple vertex-by-vertex refinement procedure [vertex mover (VM)] [16]. The essential idea of the MSG algorithm is to promote the simultaneous merging of several pairs of communities to prevent premature trapping in a local optimum of modularity. Given an appropriate choice of the step width l, the MSG-VM algorithm finds partitions with high modularity in short running time. Our implementation of the MSG-VM algorithm [16,17] has the same scaling behavior as the efficient version of the greedy algorithm [18], which has the smallest complexity among the commonly used community-detection algorithms [19]. Note that the running time of both the MSG-VM algorithm [16] and the greedy algorithm [18] is $O(DL \log N)$ with L, N, and D the number of edges, vertices, and the depth of the dendrogram describing the community structure, respectively. For a sparse network with $L \sim N$ and $D \sim \log N$, the scaling is essentially linear $O(N \log^2 N)$.

In this paper, we derive an empirical formula for predicting optimal l values—i.e., values of the step width that yield a modularity very close to the highest value achievable by the MSG-VM algorithm. Furthermore, for two real-world networks having each an inherent partition into substructures, we compare the community structures identified by the original greedy and the MSG-VM algorithm. These real-world examples are the metabolic network of *E. coli* and the graph of coappearing words in the titles of publications coauthored by Martin Karplus, the most cited theoretical chemist. The inherent substructures of the former are the metabolic pathways, while the inherent substructure of the network of Karplus' paper titles are the sets of words predominantly used in research subfields in theoretical and computational chemistry. These two examples illustrate that the

*schutz@bioc.uzh.ch
†FAX: +41 44 635 68 62 caflisch@bioc.uzh.ch

MSG-VM algorithm detects the underlying substructures more accurately than the original greedy algorithm.

II. METHODS

A. Multistep greedy and vertex mover algorithms

The MSG algorithm optimizes modularity by an iterative procedure in which multiple pairs of communities are merged at each iteration. This multistep approach is a significant extension with respect to the original greedy algorithm [14], in which only the pair of communities that improves modularity most is merged in each iteration. A pseudocode description of the MSG algorithm is

Initialization:
Each vertex is a community;
Calculate matrix ΔQ whose elements are the modularity changes upon merging of module pair (i,j);
Iteration:
while pair (i,j) with $\Delta Q_{ij}>0$ exists **do**
  for all triplets $(i,j,\Delta Q_{ij})$ of ΔQ, parsed w.r.t. decreasing ΔQ_{ij} and increasing (i,j)
  do
    if $\left\{ \begin{array}{c} \Delta Q_{ij} > 0 \text{ in best } l \text{ values in } \Delta Q \text{ - matrix} \\ i \text{ and } j \text{ unchanged in iteration} \end{array} \right\}$
    then
      MergeCommunities(i,j)
    end if
  end for
end while

Algorithm 1: Flowchart of the MSG procedure. Details of the efficient merge of two communities and the calculation of the modularity change matrix are presented in [16].

Note that the step width l influences the number of merged pairs (but is not necessarily identical to it); furthermore, l is kept constant during an MSG run (for more details, the reader is referred to the original publication [16]).

Applied upon convergence of the MSG algorithm the VM procedure improves modularity by "adjusting" misplaced vertices. The VM procedure parses the vertex list in ascending vertex degree and index order and checks for each vertex whether a reassignment to one of the neighboring communities yields a modularity improvement [16].

B. Networks

All networks in this article are treated undirected and unweighted.

1. Real-world networks

The real-world networks are the same as in [16] and are listed in Table I. Sociological applications are included with the Zachary karate club example [20], the conference graph of college football teams [21], the graph of jazz groups with common musicians [2], the network of mutual trust (PGP-key signing) [27,28], the collaboration network (coauthorships in cond-mat articles) [1], and the graph of costarring actors in the IMDB database [31]. Network applications in biochemistry are covered by the graph of metabolic reactions in the nematode *Caenorhabditis elegans* [22] and the bacterium *Escherichia coli* [3] as well as two different data sets describing the protein-protein interactions in *Saccharomyces cerevisiae* (budding yeast) [24,25] with labels "PPI" and "yeast." Linguistic applications are covered by the Word Association network [29] and the graph of the coappearing words in titles of publications (co)authored by Martin Karplus [16,17] who has the third highest h factor [33] among chemists [34]. From computer science the internet routing network [26] and the graph of WWW pages [30] are included. The effects of disconnected graphs are considered by including the full network as well as its largest connected component (LCC).

2. Computer-generated networks

A total of 1100 computer-generated networks were used for an in-depth assessment of the empirical formula for the prediction of optimal values of l (Table II). The networks in $GN_{1,2,3}$ consist of 128 vertices organized in four equally sized communities [21]. The cohesion of the vertices within a module is controlled by a parameter called z_{out} which determines the average number of edges connecting vertices of different modules. To consider clearly formed and loosly coupled modules the z_{out} parameter is chosen uniformly from the second smallest to the highest value. Among the sets $GN_{1,2,3}$, the number of edges is varied to assess the effect of different values of average degree.

The remaining test cases are larger networks, which have no imposed community structure and a heterogeneous distribution of the vertex degrees and community sizes (confer Table 1 in the supplementary material [32]). A recent study, published after the submission of this work, has emphasized the importance of this heterogeneity for testing community-detection algorithms on severe benchmarks [35]. To foster a "spontaneous" formation of modules a vertex degree distribution is imposed. The network is generated by choosing a number of vertices at random (uniform distribution), assigning edge end points to each vertex according to the degree distribution and joining the edge endpoints at random. To examine the effect of different degree distributions, an exponential distribution is used for the networks in SED (small networks with exponential degree distribution) and a linear distribution is imposed on the networks in SLD and LLD. All networks in LLD have at least 300 vertices. After generation, the networks in SED, SLD, and LLD are projected onto the largest connected component and treated as unweighted.

III. RESULTS

It is helpful to recall here that L is the number of edges and l_{opt} is the value of the step width that yields the highest MSG-VM modularity (among all tested values of step width). The MSG-VM algorithm is applied on each real-world network using every integer $l<\min\{5000,L\}$. The modularity values before and after the VM application are

TABLE I. Properties of real-world networks and comparison of MSG-VM runs using l as in Eq. (2) or picked at random. The column "Q_{opt}" lists the maximal value of modularity obtained by running MSG-VM for all values of l smaller than $\min\{5000, L\}$ (where L is the number of edges). The column "Q_{pred}" lists the MSG-VM modularity obtained using Eq. (2) to determine the step width. The columns "$\langle Q_{\mathrm{rand}}\rangle$" and "$\langle Q_{\mathrm{rand}}^{l<1.5\sqrt{L}}\rangle$" 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 (details are given in Sec. VII of the supplementary material [32]). Six values of l are picked randomly for each sample because six values were used to determine Q_{pred}: the four values of l calculated by Eq. (2) and the two integers adjacent to the best of these four. Values of $\langle Q_{\mathrm{rand}}\rangle$ and $\langle Q_{\mathrm{rand}}^{l<1.5\sqrt{L}}\rangle$ higher than the corresponding Q_{pred} are in italics. The acronym LCC stands for "largest connected component."

				MSG-VM with Optimal l		MSG-VM with l from Eq. (2)	MSG-VM with Random l	
Network	Ref.	Vertices	Edges (L)	$l_{\mathrm{opt}}/\sqrt{L}$	Q_{opt}	Q_{pred}	$\langle Q_{\mathrm{rand}}\rangle$	$\langle Q_{\mathrm{rand}}^{l<1.5\sqrt{L}}\rangle$
Zachary Karate Club	[20]	34	78	0.34	0.398	0.398	0.391	0.398
Metabolic *E. coli*	[3]	443	586	0.25	0.816	0.816	0.813	0.816
College Football	[21]	115	613	0.04	0.603	0.595	0.579	*0.596*
Metabolic *C. elegans*	[22]	453	1899	4.80	0.450	0.447	0.439	0.445
Jazz	[2]	198	2742	10.81	0.4451	0.4447	*0.4451*	*0.4448*
Email	[23]	1133	5451	0.76	0.575	0.575	0.564	0.574
Yeast (PPI, LCC)	[24]	2552	7031	0.42	0.706	0.705	0.693	0.702
M. Karplus	[16,17]	1167	13423	0.79	0.316	0.311	0.306	0.311
PPI *S. cerevisiae* (LCC)	[25]	4626	14801	1.40	0.545	0.544	0.531	0.543
PPI *S. cerevisiae*	[25]	4713	14846	1.40	0.546	0.546	0.532	0.545
Internet	[26]	11174	23409	1.82	0.625	0.619	0.615	0.618
PGP-key signing	[27,28]	10680	24340	0.28	0.878	0.876	0.873	0.876
Word Association (LCC)	[29]	7204	31783	0.40	0.541	0.536	0.528	0.536
Word Association	[29]	7207	31784	0.54	0.540	0.537	0.527	0.536
Collaboration	[1]	27519	116181	0.45	0.748	0.746	0.743	0.744
WWW	[30]	325729	1117563	2.87	0.939	0.936	*0.937*	*0.937*
Actor	[31]	82583	3666738	1.27	0.543	0.536	*0.537*	*0.539*

recorded. For the computer-generated networks all integer values $l < 10\sqrt{L}$ have been tested (the $\sqrt{L}$ scaling is rationalized in the next subsection).

A. Dependence of l on network properties

The correlation between the optimal step width l_{opt} and several topological properties was calculated. The following properties or powers thereof were used: number of vertices and edges, highest degree, average degree, standard deviation of degree, average of power 1, 2, or 3 of the clustering coefficient, and average and standard deviation of the differences in clustering coefficient values or degree of linked vertices. The highest correlation was observed for $\sqrt{L}$ (0.7728, correlation coefficients of other properties are listed in the supplementary material [32]).

This empirical result is consistent with the $\sqrt{L}$ dependence of the number of communities yielding maximal modularity as recently demonstrated for one class of networks [15]. In fact, a close inspection of the MSG algorithm shows that the step width l determines the number of communities formed during the first 1%–5% of the iterations (the number of iterations is strongly dependent on the network topology). Each module in the final solution has to be nucleated as early as possible and therefore l to be chosen according to the expected number of communities.

1. Optimal prefactor for computer-generated networks

To determine the prefactor α in the $\sqrt{L}$-scaling law the computer-generated networks introduced in Sec. II B 2 are examined first. This choice is due to their multitude (1100 networks) and their lack of overlapping condensed structures [i.e., few (almost) complete subgraphs sharing vertices] as consequence of the construction principle. First, we observe that for 97 of the 1100 networks the MSG-VM modularity does not depend on l. Further, for each value of α the MSG-VM modularity is averaged over all networks of the same type $\bar{Q}_{\mathrm{MSG-VM}}(\alpha) = \frac{1}{N_S}\sum_{i\in S}\frac{Q^i_{\mathrm{MSG-VM}}(\lfloor\alpha\sqrt{L_i}\rfloor)}{\max_l(Q^i_{\mathrm{MSG-VM}}(l))}$, where S is the type of networks, N_S is the number of networks of type S, $\lfloor . \rfloor$ is the floor function, and L_i is the number of edges in network i. All α profiles peak for $0.2 < \alpha < 0.3$ and show a similar behavior (Fig. 1). The α profiles averaged over all computer-generated networks peak at $\alpha = 0.251$. [It is legitimate to consider the average because for each α the histogram of $\frac{Q^i_{\mathrm{MSG-VM}}(\lfloor\alpha\sqrt{L_i}\rfloor)}{\max_l(Q^i_{\mathrm{MSG-VM}}(l))}$ (i indexing the network realizations) follows an unimodal distribution with an additional peak at

TABLE II. Properties of computer-generated networks. The networks in the GN_i (Girvan-Newman) sets ($i=1,2,3$) possess an imposed four community structure where z_{out} controls the average number of edges connecting two different modules [21]. For the networks of type SED (small networks with exponential degree distribution), SLD (small networks with linear degree distribution), and LLD (large networks with linear degree distribution) a degree distribution has been prescribed to foster the formation of communities.

Type	No. of realizations	Vertices	Edges	Remarks
GN_1	100	128	1024	$z_{out}=3-16$
GN_2	100	128	512	$z_{out}=2-8$
GN_3	100	128	2048	$z_{out}=2-32$
SED	300	11–976	10–19247	Exp. deg. distr.
SLD	200	19–3777	43–78741	Linear deg. distr.
LLD	300	309–4278	1523–342940	Linear deg. distr.

1.0 originating from the degeneracy of l^i_{opt}.] Excluding the additional peak, the highest normalized modularities are still observed for $0.2<\alpha<0.3$. Remarkably, the degeneracy of l^i_{opt} [i.e., the number of networks with $Q^i_{\text{MSG-VM}}(\lfloor\alpha\sqrt{L_i}\rfloor) = \max_l(Q^i_{\text{MSG-VM}}(l))$] is highest for $0.18<\alpha<0.26$. A leave-N-out procedure (confer supplementary material [32] for details) provides evidence that $\alpha=0.251$ would have been (close to) optimal also for another selection of networks. The application of the MSG-VM algorithm with step width $\lfloor 0.251\sqrt{L}\rfloor$ yields 97.6% of the highest MSG-VM modularity averaging over all computer-generated networks (98% if median is calculated).

2. Comparison of empirical formula with random selection of step width

If a step width value is selected at random among $l<\min\{L,5000\}$ (all tested values), the MSG-VM algorithm is expected to yield 93.4% of the highest MSG-VM modularity on average over all computer-generated networks [the expectation value is equal to the arithmetic mean over all $Q_{\text{MSG-VM}}(l)$ values]. An in-depth analysis (details given in the supplementary material [32]) shows that $l_{\text{opt}}<1.5\sqrt{L}$ for 92.6% of all computer-generated networks. If a step width value smaller than $1.5\sqrt{L}$ is chosen at random, the expectation value of the MSG-VM modularity raises to 95.9% of its highest value (average over all computer-generated networks). Thus, the empirical formula $l=0.251\sqrt{L}$ performs 4.3% better (of a maximum of 6.6%) than a value of step width picked at random if all tested values are considered. If the reduced test set $l<1.5\sqrt{L}$ is used, the empirical formula performs 1.7% better than a value of step width picked at random (4.1% maximal improvement). More precisely, for 85.5% of the networks the MSG-VM modularity with $l=0.251\sqrt{L}$ is higher than the one with l picked at random and the average improvement for these networks is 2.4%.

To account for limited sampling the prefactor $\alpha=0.25$ is assumed to be optimal for the computer-generated networks (the prefactors 0.251 and 0.25 can be considered identical as the real to integer conversion yields the same value of l for networks with $L<10^6$).

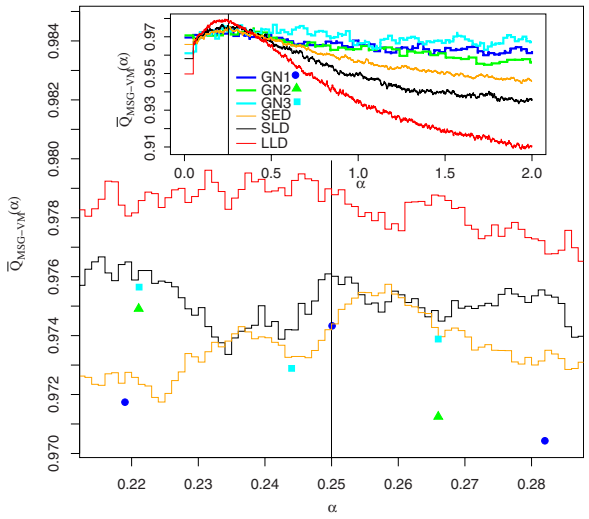

FIG. 1. (Color online) Dependence of $Q_{\text{MSG-VM}}$ on the $\sqrt{L}$ prefactor α for the computer-generated networks. The averages are taken separately for each type of computer-generated networks. The vertical black line denotes $\alpha=0.25$, which is the value suggested in Eq. (2). The parameter range for α has been discretized to multiples of 0.001 to simplify the calculations.

3. Application to real-world networks

In comparison to computer-generated graphs, real-world networks are endowed with more condensed substructures. Therefore, a different scaling behavior than for the computer-generated networks is possible. To improve statistics and reduce spurious effects due to vertex-labeling artifacts (a value of step width yields a high MSG-VM modularity as it profits exclusively from the "right" parsing of the vertices), 100 copies of the smallest 10 real-world networks are created with permuted vertex labelings (details are presented in the supplementary material [32]). For each copy the influence of l is tested as described in Sec. III. Except for the College Football and Email networks all $\bar{Q}_{\text{MSG-VM}}$ profiles (confer Sec. III A 1 for the definition) averaged over the scrambled variants are observed to peak for values of step width equal or very close to

$$l=\lfloor\alpha\sqrt{L}\rfloor \quad (\alpha=0.25,0.5,0.75,1) \tag{2}$$

(supplementary material [32]). The MSG-VM modularity deviates at most by 1.47% from the maximal value (Table I). Moreover, for 13 of the 17 networks the deviation is smaller than 0.94%. In comparison to the effect of permuted vertex labels this deviation is of the same order of magnitude and thus negligible (details given in the supplementary material [32]).

To further assess the predictive power of Eq. (2), the MSG-VM modularity obtained with l as in Eq. (2) is compared with a random selection of the step widths. Because of the real to integer conversion induced by the floor function, an integer adjacent to $\lfloor\alpha\sqrt{L}\rfloor$ might be optimal. Therefore, not only the four values of step width as in Eq. (2) are tested, but

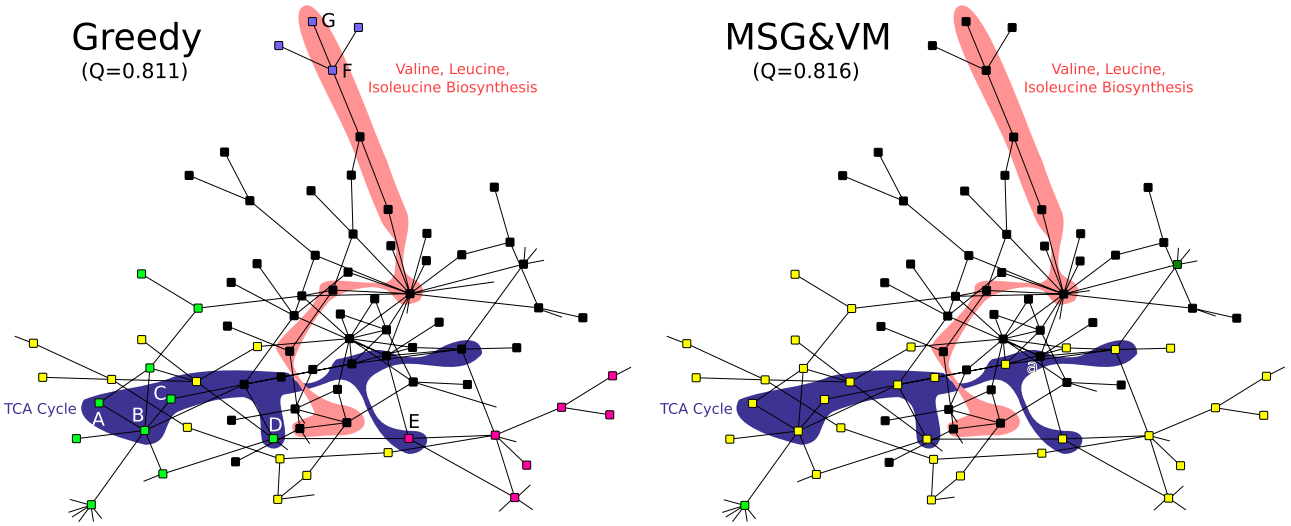

FIG. 2. (Color online) Clusterization of the metabolic network of *E. coli* and accuracy of pathway identification. Two exemplary pathways as taken from the KEGG database [36,37] (pathways MAP00020 for "TCA cycle" and MAP00290 for "Valine, Leucine, Isoleucine Biosynthesis") are highlighted by the colored areas. An excerpt of the network is shown here while the full network is in the supplementary material [32]. The misassigned vertices are indicated by letters; they are a=(S)-Malate for MSG-VM, and for the original greedy: A=3-Carboxy-1-hydroxypropyl-ThPP, B=2-Oxoglutarate, C=Oxalosuccinate, D=Succinate, E=Fumarate, F=2-Oxoisovalerate, and G =Valine.

also the two integers adjacent to the best of them. For a fair comparison the same number of trials is allowed in the random experiment. For 14 out of 17 networks the MSG-VM modularity value with l as in Eq. (2) is higher or equal than for the corresponding random experiment (Table I). Therefore, one can conclude that the empirical formula (2) yields a step width that results in (close to) optimal modularity, and therefore only six runs of the MSG-VM algorithm are required.

B. Quality of MSG-VM network partition

Previously, the performance of the MSG-VM algorithm in optimizing modularity has been shown on 19 real-world networks [16]. Here, an in-depth analysis of two examples provides evidence that the MSG-VM algorithm gathers vertices in groups that represent substructures (identified by other means) more accurately than the greedy algorithm.

1. Metabolic network of E. coli

The network of metabolic reactions in the bacterium *E. coli* is extracted from the KEGG database (data set "Escherichia coli K-12 MG1655") with additional refinement by Ma and Zeng [3] and projected on the largest connected component. Furthermore, chains of vertices with degree 1 or 2 are reduced to one single vertex (to reduce spurious effects of unnaturally splitted chains). Each vertex is assigned to between zero and eight out of 11 metabolic pathways with an average of 1.51 ± 0.99.

Eleven communities are identical in the MSG-VM partition (which consists of 19 communities and has $Q=0.816$) and the partition obtained with the greedy algorithm (20 communities, $Q=0.811$). To assess the quality of pathway detection, we employ the measure $P=\Sigma_i \frac{P_i}{N_i}$ (adopted from [5]), with P_i the number of vertex pairs in community i that share at least one pathway and N_i the number of pairs of vertices with equal community affiliation. The MSG-VM partition has $P_{\mathrm{MSG-VM}}=0.60$, which is better than the partition obtained with the original greedy algorithm ($P_{\mathrm{greedy}}=0.58$). The improved pathway identification is illustrated by an excerpt of the network in Fig. 2 (vertices in the 11 modules which are identical in both partitions are removed for visibility reasons). Two central pathways (classification according to KEGG database) are highlighted by colored areas. In the MSG-VM solution the vertices of each pathway belong to separate modules except for "(S)-Malate." This metabolite has more connections to vertices assigned to the "Amino Acid Metabolism" than to those of the "Carbohydrate Metabolism" (the "TCA cycle" is associated to the latter) and thus, a separation is meaningful. On the other hand, the metabolites misclassified by the original greedy algorithm are "2-Oxo-glutarate" (B), "3-Carboxy-hydroxypropyl-ThPP"(A), and "Oxalosuccinate"(C). The last two belong only to the "TCA cycle" pathway, whereas "2-Oxo-glutarate" is part of several pathways and therefore can also be attributed to other communities. Furthermore, the separation of the blue vertices in the "Valine, Leucine, Isoleucine Biosynthesis" pathway is peculiar as the overlapping pathway "pantothenate and CoA biosynthesis" is contracted to one vertex (the vertex right to "F" and "G"). The metabolites "F" and "G" are the educts in the "pantothenate and CoA biosynthesis" pathway. If a unique assignment has to be made, an attribution to the "Valine, Leucine, Isoleucine Biosynthesis" pathway is more reasonable. The last differences of the greedy partition to the MSG-VM solution are "Succinate" (D) and "Fumarate" (E) which are as "(S)-Malate" (a) part of multiple different metabolic processes and therefore may be attributed to multiple pathways. To summarize, of eight assignments differing between MSG-VM and original greedy

TABLE III. The five largest communities as identified by the MSG-VM algorithm in the network of words in the titles of M. Karplus' papers. These five communities account for 81% of the vertices in the network. Unspecific words (e.g., "study" and "theory" with degree 291 and 234, respectively) were taken into account for the clusterization, but are not listed in this table.

		Most frequent words		Number of titles with any of the words in community	
Rank	Vertices	Degree	Word		Description
1	220	407	Protein	442	Molecular dynamics (of proteins)
		318	Simulation		
		269	Molecular-dynamics		
2	184	290	Structure	330	Three-dimensional structures
		123	Peptide		
		97	Inhibitor		
3	162	269	Model	335	Molecular modelling, molecular mechanics
		178	Energy		
		169	Function		
4	162	159	Molecule	306	Quantum mechanics, free-energy calculation
		154	Free-energy		
		144	Potential		
5	116	212	Reaction	205	Chemical reaction, kinetics, and solvation
		154	Solution		
		101	Solvation		

algorithm (in the excerpt of the network shown in Fig. 2), none was misplaced by the MSG-VM algorithm, whereas the greedy algorithm misplaced two metabolites (two further examples of incomplete detection of pathways by the original greedy algorithm are shown in the supplementary material [32]).

2. Network of words in titles of Karplus' publications

Martin Karplus is one of the most productive and most cited chemists (78091 citations as of 3 July 2008). As second example we analyze the community structure of the graph of words coappearing in the titles of the 719 publications coauthored by Karplus between 1947 and 2004 [16,17]. The words with highest degree in the five largest (according to number of words) communities are shown in Table III.

The following two examples provide evidence for the superiority of the MSG-VM partition (11 communities, $Q=0.316$) with respect to the partition obtained by the original greedy algorithm (18 communities, $Q=0.264$). The words "reaction" (degree 212), "hydrolysis" (73), "rate" (69), "enzyme" (57), "catalysis" (54), and "kinetics" (54) are appropriately grouped in a single community by the former, while they are spread in the four largest (according to the number of words) communities by the latter. Another example of superiority of the MSG-VM partition is the community with the words "molecule" (159), "atom" (91), and "bond" (87), which are spread over the three largest communities by the greedy algorithm. These two examples show that the main advantage of the MSG-VM algorithm is that the simultaneous emergence of several communities hinders the spurious coalescence into few large communities observed for the original greedy algorithm.

IV. CONCLUSIONS

The performance of the MSG procedure, a multistep extension of the greedy algorithm, was analyzed on 1100 computer-generated networks of heterogeneous sizes and degree distributions and 17 real-world networks. Several powers of topological properties (e.g., average degree, clustering coefficient, etc.) were tested as prediction formulas for the optimal step width l. The empirical formula $l=\lfloor\alpha\sqrt{L}\rfloor$ (L total edge weight; $\alpha=0.25,0.5,0.75,1$) outperforms all others and yields a higher modularity value than a random picking of the step width for 85.5% of the computer-generated networks and 14 of 17 real-world examples. For these 14 real-world networks, the modularity optimized by MSG-VM algorithm using only six values of l ($l_1=\lfloor 0.25\sqrt{L}\rfloor$, $l_2=\lfloor 0.5\sqrt{L}\rfloor$, $l_3=\lfloor 0.75\sqrt{L}\rfloor$, $l_4=\lfloor 1.0\sqrt{L}\rfloor$, and $l_{5,6}=l_{\max}\pm 1$ with $l_{\max}$ the step width among $l_{1,\ldots,4}$ that yields the highest modularity) is larger than 99% of the highest value achievable by exhaustive testing of all step widths (i.e., $1\leqslant l\leqslant L$). This deviation is on the order of the fluctuations observed when the parsing order of the vertices is changed. In addition, for 92.6% of the computer-generated and 13 of 17 real-world networks the optimal value of the step width is smaller than $1.5\sqrt{L}$.

To assess the quality of the community identification two real-world examples (the network of metabolic reactions in *E. coli* and the graph of coappearing words in titles of publications coauthored by M. Karplus) were examined in-depth and the modular structure obtained from the application of the MSG-VM and greedy algorithms was compared. For the metabolic network the original greedy algorithm splits two exemplary pathways ("TCA cycle" and "Valine, Leucine, Isoleucine Biosynthesis") in multiple parts with seven misplaced vertices. Two of these vertices are not part of another pathway and therefore are wrongly assigned by the original greedy algorithm. For the MSG-VM solution only one metabolite is misplaced which can be attributed to the three pathways in which this metabolite is involved. Furthermore, an objective criterion (the conditional probability that two vertices in the same module share at least one pathway) sup-

ports these exemplary observtions. For the "M. Karplus" network the partition obtained by the original greedy algorithm has three very large modules in which words of distinct research fields are inappropriately mixed. Moreover, subsets of words belonging to the same topic are erroneously split (e.g., "atom," "molecule," and "bond" are split in the three largest modules). On the other hand, the MSG-VM procedure more accurately groups subsets of words belonging to individual research topics.

In conclusion, the MSG-VM algorithm is one of the fastest and most accurate procedures for modularity optimization currently available because it scales as $O(N \log^2 N)$ for a sparse network (N the number of vertices) [16]. Therefore, a single run is faster than previously published approaches [19], and only six independent runs are required using Eq. (2) to determine the step width [17].

ACKNOWLEDGMENTS

We thank Stefanie Muff for helpful comments on the manuscript. Christian Bolliger, Thorsten Steenbock, and Dr. Alexander Godknecht are acknowledged for maintaining the Matterhorn cluster where most of the presented parameter studies were performed. For providing the data sets we are thankful to Dr. A. Arenas, Dr. A. L. Barabasi, Dr. P. Gleiser, Dr. H. Ma, and Dr. M. E. J. Newman [17]. This work was supported by a Swiss National Science Foundation grant to A.C.

[1] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[2] P. Gleiser and L. Danon, Adv. Complex Syst. **6**, 565 (2003).
[3] H. Ma and A.-P. Zeng, Bioinformatics **19**, 270 (2003).
[4] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[5] S. Muff, F. Rao, and A. Caflisch, Phys. Rev. E **72**, 056107 (2005).
[6] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).
[7] I. A. Hubner, E. J. Deeds, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **102**, 18914 (2005).
[8] D. Gfeller, P. D. L. Rios, A. Caflisch, and F. Rao, Proc. Natl. Acad. Sci. U.S.A. **104**, 1817 (2007).
[9] S. Muff and A. Caflisch, Proteins **70**, 1185 (2008).
[10] S. Krivov, S. Muff, A. Caflisch, and M. Karplus, J. Phys. Chem. B **112**, 8701 (2008).
[11] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[12] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner, e-print arXiv:physics/0608255.
[13] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[14] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[15] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).
[16] P. Schuetz and A. Caflisch, Phys. Rev. E **77**, 046112 (2008).
[17] The source code of the MSG-VM algorithms, the awk scripts to generate the SED, SLD, and LLD networks, as well as the network of words in the titles of Martin Karplus' publications are available at http://www.biochem-caflisch.uzh.ch/communitydetection
[18] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[19] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, J. Stat. Mech.: Theory Exp. 2005, P09008
[20] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1974).
[21] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[22] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, Nature (London) **407**, 651 (2000).
[23] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, Phys. Rev. E **68**, 065103(R) (2003).
[24] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, *et al.*, Nature (London) **440**, 637 (2006).
[25] V. Colizza, A. Flammini, A. Maritan, and A. Vespignani, Physica A **352**, 1 (2005).
[26] Internet Network: undirected, unweighted network of the Internet at the Autonomous System level from data collected by the Oregon Route Views Project (http://www.routeviews.org/) in May 2001, where vertices represent Internet service providers and edges connections among them. The file reports the list of connected pairs of nodes.
[27] X. Guardiola, R. Guimerà, A. Arenas, A. Díaz-Guilera, D. Streib, and L. A. N. Amaral, e-print arXiv:cond-mat/0206240.
[28] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, Phys. Rev. E **70**, 056122 (2004).
[29] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, Behav. Res. Methods Instrum. Comput. **36**, 402 (2004).
[30] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).
[31] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[32] See EPAPS Document No. E-PLEEE8-78-103808 for supplementary material, which includes the in-depth characterisation of the computer-generated networks, the stability analysis of Eq. (2), the explanation for the degeneracy of l, details on the calculation of Q_{pred}, correlation coefficients of optimal values of l and topological properties of the network, and a representation of the metabolic network of *E. coli*.. For more information on EPAPS, see http://www.aip.org/pubserv/epaps.html.
[33] J. E. Hirsch, Proc. Natl. Acad. Sci. U.S.A. **102**, 16569 (2005).
[34] P. Ball, Nature (London) **448**, 737 (2007).
[35] A. Lancichinetti, S. Fortunato, and F. Radicchi, e-print arXiv:0805.4770.
[36] M. Kanehisa and S. Goto, Nucleic Acids Res. **28**, 27 (2000).
[37] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, Nucleic Acids Res. **30**, 42 (2002).