Network motifs in computational graphs: A case study in software architecture

Sergi Valverde\textsuperscript{1} and Ricard V. Solé\textsuperscript{1,2}

\textsuperscript{1}ICREA-Complex Systems Lab, Universitat Pompeu Fabra, Dr. Aiguader 80, 08003 Barcelona, Spain
\textsuperscript{2}Santa Fe Institute, 1399 Hyde Park Road, New Mexico 87501, USA

(Received 1 July 2004; revised manuscript received 3 June 2005; published 8 August 2005)

Complex networks in both nature and technology have been shown to display characteristic, small subgraphs (so-called motifs) which appear to be related to their underlying functionality. All these networks share a common trait: they manipulate information at different scales in order to perform some kind of computation.

Here we analyze a large set of software class diagrams and show that several highly frequent network motifs appear to be a consequence of network heterogeneity and size, thus suggesting a somewhat less relevant role of functionality. However, by using a simple model of network growth by duplication and rewiring, it is shown the rules of graph evolution seem to be largely responsible for the observed motif distribution.

\textbf{I. INTRODUCTION}

Many natural and artificial systems are describable as networks of interacting components [1–4]. The network is a medium that allows resource sharing often involving an efficient transport of energy (metabolism, power grid), matter (highways, airport webs), or information (cellular communication, Internet). The architecture of complex networks can be explored at different scales, from the overall properties defined by average measures such as path length or clustering, correlations, or degree distributions to the more fundamental features displayed by small subsystems. In this context, it has been shown that some special, small subgraphs—so-called motifs—seem to be particularly relevant in describing the architecture of complex networks [5]. Motifs have been suggested to be the functional building blocks of network complexity. Are some subgraphs more common than others because their functional relevance?

An alternative view is that the rules of network growth can by themselves favor some subgraphs with no special relation to the underlying functionality. Actually, this seems to be the case for the structure of the protein-protein interaction map. In spite of the fact that proteins perform functions, the overall architecture of the protein network is easily reproduced by means of a simple model of node duplication plus rewiring [6]. Such properties include scale freeness, small-world features, and even hierarchical organization and protomodularity [7]. Mounting evidence suggests that many key features of complex networks (including motifs) might be strongly tied to the global network structure [8].

If functional constraints to network architecture have to be considered, one particularly relevant aspect of network complexity is associated with the presence of some underlying computational process. Computation is a key ingredient of any complex adaptive system (CAS). By storing and processing information, CAS’s are able to predict (and adapt to) external fluctuations. Computation occurs in both natural and artificial systems [9], although the building process that creates the computational structure is different. This is actually one of the most important points here: are the rules of designed and evolved systems completely different? Biological networks are largely originated through tinkering [10–12]: new components are obtained by re-using old ones, mainly by duplicating them. In spite of the apparent limitations of such mechanisms, it allows one to discover good designs [13]. More complex computations can be developed as the network size is increased and new functions can emerge.

How is computation linked to network structure? Tentative answers, to be developed here, can be obtained by looking at a very important class of computationally driven networks: software systems. They offer a unique opportunity of exploring different levels of complexity with well-defined functional traits. As opposed to most examples of evolving computational networks, extensive databases storing software evolution registers exist and involve a high degree of detail. Here we analyze the largest data set of software maps explored to date (83 different systems). The main goal of our study is to see if functionality, as opposed to network evolution, is a main constraint to the distribution of network motifs in real graphs. The paper is organized as follows. In Sec. II an overview of the software systems analyzed here is given. In Sec. III, the statistical patterns of network motifs in a large set of software systems are presented and the presence of scaling relations and the size-dependent frequency of motifs analyzed. In Sec. IV a model of duplication and rewiring is used in order to reproduce the structure of motifs of a large software map. In Sec. V a general discussion is presented.

\textbf{II. SOFTWARE NETWORKS}

Programming languages describe software systems [14]. Every computer program has a textual representation following syntactic rules dictated by a programming language [15]. The program is decomposed in a number of simpler software entities or logical elements, which are given a unique name. Software entities include things as data objects, instructions, subprograms, or modules. A hierarchy or natural ordering between software entities is prompted by modern programming languages. At the lowest level, a program is viewed as a sequence of simple machine instructions. Sequences of related instructions are enclosed in subprograms. At the highest level, there are modules or logical containers grouping simpler software entities. Often, modules are defined as functional blocks but there are no strict principles driving module composition.
It is useful to depict the complex structures defined in computer programs by means of a graph [16], where nodes represent software entities and links represent syntactical relationships between modules, subprograms, and instructions (see Fig. 1). In this paper, we will focus on a particularly interesting subset of software entities: the collection of modules and their static interactions (also called software architecture). While it is widely acknowledged that software evolution depends on its architecture, very little is known about the cause and effect relationships between design practices and evolution outcomes [17]. In order to understand the relation between structure and artificial evolution, we have envisaged a network model of software architecture, hereafter called the software map or software network. Here, we show how the evolution of computer programs can be understood by recovering and analyzing their software networks at different stages of development.

Following [16,18], Fig. 2 shows the text of an incomplete C++ program [see Fig. 2(a)] and its corresponding software map [see Fig. 2(b)]. The program text reads from left to right and top to bottom. The software network \( \Omega = (V, L) \) of this C++ program is recovered by means of a very simple lexical analysis. First, we identify the vector of all module names (also class in C++) given by \( W = (w_i) \) (point, chessmen, point, move, point, point, pawn, chessmen, move). Name ordering is important when recovering module dependences (see below). Names \( w_i \) that appear in the head of a module declaration provide the set of network nodes \( V = \{v_i\} \). These names (hereafter called module definitions) are easily identified because they are proceeded by the C++ keyword \textit{class}. Remaining names are called module references. In this example, we have four \( (N = \vert V \vert = 4) \) unique module definitions: point \((w_i)\), chessmen \((w_2)\), move \((w_4)\), and pawn \((w_7)\). This defines a mapping from names \( w_k \in W \) and network nodes \( v_i \in V \) in the software network.

The design of any nontrivial function involves the interaction of at least two modules [20]. Static module interactions can be depicted from relative positions of names in \( W \). Let us assume that \( w_k \in \{w_1, w_2, w_4, w_7\} \) is a module definition associated with node \( v_j \) and \( w_k \notin \{w_1, w_2, w_4, w_7\} \) is a module reference associated with node \( v_j \). A directed link \( \{v_j \rightarrow v_i\} \in L \) signals a dependence from module definition \( w_k \) to module reference \( w_l \). Link directionality reflects name ordering in the C++ program—that is, \( k < l \). There are two types of module dependences: association (also "has a" relationship) or inheritance (or "is a" relationship). The purpose of these dependences is to establish a logical organization of the system. However, our analysis is centered on the study of topological patterns and does not take into account detailed relationship semantics. In an association, referenced node \( v_j \) is nested in the C++ block of module \( v_i \). This block is always bracketed by the symbols [and]. In an inheritance, referenced node \( v_j \) always follows the C++ sequence: public after the referencing node \( v_i \) [see Fig. 2(a)]. Repeated links are not considered in the following analysis.

Software maps capture the topology of complex software systems. In particular, these maps provide a quantitative approach to the evolution of technology. They are actually evolving entities and somewhat inhabit an intermediate zone between computing machines and neural structures. We have shown software networks to be scale free and small world [16,18,21]. Software networks can be described under a statistical physics perspective.

III. SOFTWARE MOTIFS

In this section, we extend our previous topological studies by analyzing software networks at the level of network subgraphs, or subsets of connected nodes in a network. The statistics of subgraphs provides important information about network structure. It has been claimed that overrepresented subgraphs (i.e., motifs) signal key building blocks of networks [5]. This might be the case for regulatory networks, where specific subgraphs (i.e., feed-forward loops) perform information processing functions [22]. A particular class of subgraphs, cycles, have received considerable interest. Cyclical dependences in software maps imply that a module is related to itself, which may be acceptable, unacceptable, or required [23]. Ambiguity in the functional meaning of cycles suggests that subgraphs in software graphs are not strictly related to well-defined functions. The ubiquity of subgraphs in software networks seems to be a consequence of top-down mechanisms of software organization and not a consequence of selective pressures.

Following the method outlined in previous section, we have recovered and analyzed a large dataset of software

![FIG. 1. Examples of common network motifs with \( n=4 \) elements found in software graphs. Here each node is a class and arrows indicate static dependences among classes (see text).](image)

![FIG. 2. (a) A piece of C++ code from a chess-playing program and (b) the corresponding software map or network model displaying the collection of modules and their logical dependences. The only information required to recover the software map is the set of module names [highlighted in bold in (a)] and their relative locations in the C++ program (see text). Notice how nodes are labeled with their names and links are decorated with relationship type.](image)
maps \{\Omega\} from 83 reverse-engineered C++ software systems. A given graph can be characterized by a degree sequence. For the whole graph \(\Omega\) each node has a degree sequence given by the in-degree list \(K_i\) and the out-degree list \(R_i\) with \(i = 1, \ldots, N\). The lists would be completed by the so-called mutual edges \(M_i\)—i.e., cases where there is a pair of edges in both directions between two nodes. For each subgraph \(\Omega_i\) of size \(n\) (here \(n=3,4\)), another degree sequence would be provided by two new lists, now \(k_{j_i}\) and \(r_{j_i}\) for the in- and out-degrees, respectively. For example, for the left subgraph in Fig. 1, we would have \(\{k_i\} = \{0,1,1,2\}\) and \(\{r_i\} = \{2,1,1,0\}\).

Network motifs are defined in terms of subgraphs which appear much more often than expected from pure chance. Specifically, they occur with a high frequency compared with the expected from an ensemble of randomized graphs with identical degree structure [5]. For each motif, we display the number of occurrences in the network \(N_{\text{real}}\), the number of occurrences \(N_{\text{rand}} \pm \text{SD}\) in a set of 100 randomized network versions, and a qualitative measure of its statistical significance as given by the \(Z\) score (see text). Medium and large software networks share a large amount of motifs but we found larger variability in the medium data set. A remarkable difference is motif 2190 (the last motif in the second row), which appeared only in the context of large software systems.

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
<th>(N_{\text{real}})</th>
<th>(N_{\text{rand}})</th>
<th>(Z_{\text{score}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software Networks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(medium)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faime</td>
<td>106</td>
<td>180</td>
<td>41</td>
<td>11</td>
<td>8.94</td>
</tr>
<tr>
<td>Aime</td>
<td>143</td>
<td>319</td>
<td>77</td>
<td>29</td>
<td>7.86</td>
</tr>
<tr>
<td>Filementa221</td>
<td>186</td>
<td>331</td>
<td>54</td>
<td>26</td>
<td>4.38</td>
</tr>
<tr>
<td>Arctic</td>
<td>255</td>
<td>391</td>
<td>68</td>
<td>26</td>
<td>3.72</td>
</tr>
<tr>
<td>Excel</td>
<td>261</td>
<td>404</td>
<td>90</td>
<td>56</td>
<td>6.01</td>
</tr>
<tr>
<td>Software Networks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(large)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>blender226</td>
<td>495</td>
<td>834</td>
<td>486</td>
<td>138</td>
<td>11.4</td>
</tr>
<tr>
<td>gg221</td>
<td>748</td>
<td>1147</td>
<td>644</td>
<td>189</td>
<td>13.4</td>
</tr>
<tr>
<td>vdk</td>
<td>771</td>
<td>1362</td>
<td>512</td>
<td>263</td>
<td>6.27</td>
</tr>
<tr>
<td>jv2</td>
<td>1364</td>
<td>1947</td>
<td>816</td>
<td>189</td>
<td>15.7</td>
</tr>
<tr>
<td>pcorely</td>
<td>1993</td>
<td>4987</td>
<td>22790</td>
<td>1840</td>
<td>1.22</td>
</tr>
<tr>
<td>Gene Regulation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(transcription)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. coli</td>
<td>424</td>
<td>519</td>
<td>40</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Neurons</td>
<td>252</td>
<td>509</td>
<td>125</td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>C. elegans</td>
<td>252</td>
<td>509</td>
<td>125</td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>Electronic Circuits</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(forward logic chips)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s15850</td>
<td>10383</td>
<td>14240</td>
<td>424</td>
<td>262</td>
<td>285</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIG. 3. Network motifs with \(n=3,4\) elements found in software graphs. The numbers of node and edges for each network are shown. The most frequent motifs were classified in distinct rows for each type of system: medium software systems, large software systems, gene regulatory nets, neural networks, and digital electronic circuits. For each motif, we display the number of occurrences in the network \(N_{\text{real}}\), the number of occurrences \(N_{\text{rand}} \pm \text{SD}\) in a set of 100 randomized network versions, and a qualitative measure of its statistical significance as given by the \(Z\) score (see text). Medium and large software networks share a large amount of motifs but we found larger variability in the medium data set. A remarkable difference is motif 2190 (the last motif in the second row), which appeared only in the context of large software systems.
tion. This is the case of Bi-parallel (S904), Bi-fan (S204), the feed-forward loop (S38), and its close variants (such as S2186 and S408). Such a common point might be easily interpreted in functional terms: similar subgraphs are abundant because they are selected or chosen to perform a given function or task. As shown below, no evidence from statistical patterns supports such view.

Assuming sparse graphs \((K) \ll N\), the probability of a given subgraph \(\Omega\) can be estimated. Following Itzkovitz et al., [24] we can see how this is calculated using the first subgraph in Fig. 1. Here we have \(\{K_{i}\} = \{2, 1, 1, 0\}\) and \(\{R_{j}\} = \{0, 1, 1, 2\}\). The idea is to compute the different probabilities associated with each directed edge linking all pairs of nodes. For example, the probability of having a directed link from node 1 to node 2 (for \(K_{i}R \leq N(K)\)) is approximately

\[
P(1 \to 2) = \frac{K_{1}R_{2}}{N(K)},
\]

which can be interpreted as follows [24]: we perform \(K_{i}\) attempts for the first node to connect to the target node with a probability \(R_{2}/N(K)\). Similarly, we would have

\[
P(1 \to 3) = \frac{(K_{1} - 1)R_{2}}{N(K)}
\]

being the approach used for all edges. The average number of appearances of \(\Omega\) is finally computed by averaging. Itzkovitz et al. [24] have shown that the average number of appearances \(\langle G \rangle\) of a given subgraph is given by a product of moments of different orders of the in-degree, out-degree, and mutual degree distributions:

\[
\langle G \rangle \sim N^{g_{a} - \gamma_{m}(K)g_{m}(M)}\sum_{j=1}^{n} \left( \frac{K_{i}}{k_{j}} \right)^{j} \left( \frac{R_{j}}{r_{j}} \right) \left( \frac{M_{j}}{m_{j}} \right),
\]

where \(g_{a}\) and \(g_{m}\) are the number of single and mutual edges. The approximation assumes uncorrelated, sparse networks \((K) \ll N\). Both conditions are met by software maps [25]. These mean-field quantities can be used as a null model estimate of the number of motifs and, eventually, to detect stray, significant deviations form randomness. Since different motifs are found in different systems [5], they can actually allow us to identify the basic functional blocks for a given class of networks.

By exploring our collection of software graphs, we determined \(\langle G \rangle\) for real nets (indicated as \(N_{\text{real}}\) in Fig. 2) and compared them to \(N_{\text{rand}}\). Here software maps with a size \(N > 10\) have been analyzed. Two groups have been chosen, involving medium-sized graphs \((N < 300)\) and large graphs \((N > 300)\). The previous set is compared with results from other networks involving computational tasks. Here previous results for both gene and neural networks are also shown for comparison (data from [5]). The reason for using biological networks as a reference system is twofold. First, the chosen systems are known to perform computational tasks (or can be described by means of an equivalent computational circuit). The second is that it has been conjectured that both natural and artificial networks might share some commonalities relating the mechanisms that shape their evolution [12]. Common features might reflect common functional traits, but also (as shown below) common rules of graph evolution with no special functional meaning.

In order to explore the question of how relevant the overall network structure is in conditioning the frequency of given subgraphs, we should consider the global structure of the network. The first approximation is to consider the degree of heterogeneity as provided by the distribution of links. Software systems have a well-defined scale-free indegree distribution

\[
P_{i}(k) = \frac{\gamma_{i} - 1}{k_{0}^{\gamma_{i}}} (k + k_{0})^{-\gamma_{i}},
\]

with \(\gamma_{i} \approx 2\). A mean value \(\langle \gamma_{i} \rangle = 2.09\pm0.05\) has been obtained by averaging over all the systems studied here. The distribution of scaling exponents is strongly peaked around \(\gamma_{i} = 2\). The out-degree distribution \(P_{o}(k)\) is much steeper and seems better described by a broad scale distribution, not far from the exponential limit. This is actually the opposite situation considered in [24] but is not difficult to show that it is essentially symmetric in the theoretical treatment.

For the regime considered here, it was shown that \(\langle G \rangle\) follows a scaling law

\[
\langle G \rangle \sim N^{\alpha}.
\]

Specifically, for a given pair \(\{n, g\}\) and a given scaling exponent, we have

\[
\langle G \rangle \sim N^{\alpha_{s} - s \gamma_{m} + \gamma_{i} + 1},
\]

where \(s\) is the maximum indegree for our case. This scaling is actually valid for \(2 < \gamma_{i} < s + 1\).

Four examples of the observed scaling laws are shown in Fig. 4 for different software motifs. Using Eq. (3) the expected number of times a given motif appears would scale as \(\langle G \rangle \sim N^{\alpha_{s} - \gamma_{i}}\) and using the scaling exponent \(\gamma_{i} = 2.09\pm0.06\) a scaling law \(\langle G \rangle \sim N^{\alpha_{s}}\) would be predicted for uncorrelated, sparse graphs. Here we use our set of systems\(^1\) whose size will be indicated as \(n_{i} (i=1, \ldots, 83)\)\) (number of nodes of each graph). For convenience, we order the systems by increasing size (i.e., \(n_{i} < n_{i+1}\)). If \(G(n_{i})\) is the number of times a given subgraph appears in the \(i\)th system (of size \(n_{i}\)), we should expect a scaling relation \(G(n_{i}) \sim n_{i}^{s}\). In order to reduce the noise level we will use the cumulative distribution

\[
G_{\text{cum}}(N) = \sum_{n_{i} < N} G(n_{i}).
\]

The cumulative distribution should scale as \(G \sim N^{\alpha_{c}}\) with \(\alpha_{c} \approx \alpha + 1\). As shown in Fig. 4, the predicted scaling is recovered from real data, thus indicating that the average trends are consistent with the expectation from random scale-free networks. It confirms the validity of the prediction of Itzkovitz et al. [24] and its agreement with a set of real networks. This agreement is an interesting result, particularly if

\(^1\)Although a total of 83 systems have been used, the presence of a specific subgraph is size dependent. Not all systems exhibit all subgraphs: for small software maps some subgraphs are absent.
network motifs in computational graphs: a

1.86±0.16, and S2186, with n=4, g=4, and s=3. The predicted exponents (using the average scaling exponent for the in-degree distribution γ=2.1) would be α(S904)=0.9, α(S472)=α(S206)=0.1, and α(S2186)=2.1, respectively. Using the cumulative number of graphs, Gcum (see text), we obtain αS(S904)=1.86±0.16, αS(S472)=0.97±0.07, αS(S206)=1.18±0.17, and αS(S2186)=3.12±0.11, in good agreement with the predicted values. The fit was made using least squares on a log-log scale.

we remember that this is a designed system to perform given functions. The fact that we obtain the scaling law expected for the random, scale-free graph reveals that the observed scaling in motif abundances are a consequence of top-down constraints derived from graph evolution.

IV. DUPLICATION-BASED EVOLUTION

The topology of software architecture emerges from designed evolution. On top of the process, there must be a basic building plan towards a final function or set of functions. The engineer foresees the outcome of its work. But there are a number of strong constraints no less important and operating through the software building process. On the one hand, modular structures are shaped through parallel paths of evolution. Different blocks will be involved in more specific subfunctions. On the other hand, increased complexity leads to conflicts between different subparts. This is reflected, for example, at the topological level: small software maps tend to display tree structure, whereas larger systems typically display much more complex patterns [16]. The common overall structure detected in software graphs in terms of the degree distribution (and other average properties) suggests that the final topological patterns might be strongly constrained.

We conjecture that the abundance of subgraphs in software networks relates to universal mechanisms of network growth underlying their evolution. Real software maps tend to display motif generalizations or subgraphs having a structure comprising many replicas of the four motifs observed here. These structures are highly redundant. This suggests a very simple duplication-based mechanism of subgraph generation. New modules depend upon other modules in order to provide useful functionalities. And it seems reasonable to assume that similar modules will share a large number of module dependences. In a related software engineering study [26], structural similarities in C++ software at the module (class) level have been analyzed. They have found quantitative evidence of structural duplications. However, they did not provide any model explaining the origin of duplications.

Figure 5 shows a detailed example illustrating how top-down duplication works in software development. Imagine we want to add a new software module representing the queen, in the previous chess-playing program [see Fig. 2(a)]. First, we will add a new module declaration, which is conveniently named queen [see Fig. 5(a)]. Because a queen is a type of chessman, it seems reasonable to make this module depend upon the same modules referenced by similar modules, which in this case is the pawn. By using the pawn module and its neighborhood as a template, we add an inheritance relationship from the queen to the chessman [see Fig. 5(b)]. Duplication is completed with the addition of a collaboration relationship from queen to move [see Fig. 5(c)]. Comparison between final network [see Fig. 5(c)] with the initial network (see Fig. 2) reveals a new biparallel subgraph and twice the number of bifan subgraphs.
An example of this process taking place in real software development is shown in Fig. 6. Here a given subsystem inside a growing software graph is displayed at different development stages. Duplication of nodes seems to be at work, as well as further removal of many links associated with a given hub. From c to d a duplication of the hub involving many incoming links has taken place (together with some further node addition). From d through i, it is evident how a large number of new classes were added by copying the pattern of single nodes connected to two central hubs. There is extensive rewiring in some stages, such as in f, where the lower hub loses a large fraction of in-links. Moreover, there is also the addition of new connections between existing nodes (see h→i). The whole sequence spans 1 year of development. The main observation from this example (which is a typical one) is that node duplication plus rewiring, particularly link removal, is widespread. This is also the case in the evolution of cellular networks [6].

Examples like the previous one suggest that duplication-divergence growth is the cause of the observed subgraph abundances in software maps. This hypothesis can be tested by comparing the distribution of subgraphs in real networks with those obtained with a stochastic model of network growth based on asymmetric duplication-divergence rules, previously described in [6]. First, an initial random (or backbone) network of \( m_0 < N \) nodes is created. This random graph is generated by the addition of nodes with degree \( k_0 = 2 \), every link pointing to a random target node [27]. This backbone posses a treelike structure (as occurs with software maps at the beginning of their evolution). Starting from this backbone, we apply the following rules at each iteration of the model.

(i) **Duplication.** A randomly chosen target node \( v \) is cloned, and the new node \( w \) attaches to all the neighbors of the target node.

(ii) **Divergence.** For each pair of original and redundant links remove one of them with probability \( \delta \).

(iii) **Cross linking.** In addition, the target and new node are linked (\( w \rightarrow v \)) with probability \( \beta \). This rule is important in order to generate triads or 3-subgraphs.

In spite of the simple set of rules implicit in the duplication model, the frequencies of subgraphs obtained from our *in silico* system are remarkably close to those seen in their real counterparts. In Fig. 7, we have compared the concentration of 4-subgraphs expressed in various software networks and the concentration of 4-subgraphs predicted with the duplication-based evolution model. These plots were obtained with the following method. We generate 400 graphs, 100 for each of four different software networks: Blender, Filezilla, GTK, and Exult [28]. Each synthetic graph has the same number of links \( L \) and number of nodes, \( N \), as measured in the corresponding software map and no further constraints are imposed. The parameter space is sampled uniformly. Once the synthetic networks are obtained, we perform a 4-subgraph census by counting the number of appearances of each 4-subgraph \( \Omega_i \) in the model and in the synthetic network. Notice that we do not test for statistical significance (as in the motif analysis). Instead, our comparison test is based solely on raw subgraph counts. In order to compare the two systems, the raw number of subgraphs of size 4 is computed and the concentration \( C \) of subgraphs evaluated. Here, the concentration is simply the number of appearances of the 4-subgraph over the total number of 4-subgraphs found.

In Fig. 7, each point represents the pair \( (C_{\text{observed}}, C_{\text{predicted}}) \) of observed and predicted concentrations for given 4-subgraph \( \Omega_i \). Specifically, we display the set of

![Figure 6](image-url)  
**FIG. 6.** A real instance of software network growth from a well-defined subsystem of Prorally [16] showing duplication. Evolution goes from top to bottom and left to right. Only the largest connected component is displayed here. The figure shows how the target hub in (c) has been duplicated in (d) (both nodes highlighted with a dotted box). Many duplicated nodes involve less connected targets [see (g) and (h)].

![Figure 7](image-url)  
**FIG. 7.** Comparison of observed and predicted (from a duplication model) 4-motif concentrations for (a) Blender, (b) Filezilla, (c) GTK, and (d) Exult (here concentrations are rescaled by \( \times 10^{-3} \)). The exponents for the least-squares fit are (a) \( \xi = 0.94 \pm 0.12 \), (b) \( \xi = 0.92 \pm 0.13 \), (c) \( \xi = 0.96 \pm 0.11 \), and (d) \( \xi = 1.14 \pm 0.12 \), respectively.
motifs: the value of the exponent 
ably well the concentration of common software network
the simple duplication model presented here predicts reason-
we restrict ourselves to
the selected model graphs to be compared with the real ones.
uncorrelated character of software maps.
This result is supported by previous observations of the
motions. This result is supported by previous observations of the
in software graphs.
Natural

by using recent theoretical and numerical methods to
In this paper we have analyzed the statistical patterns of
are of fundamental importance. In spite of such constraints, it
sign from scratch choosing sets of small gates and building optimal,
re-use is at work.
Although sharing common motifs seems to call for common
=1.06±0.18, respectively. Consistently with previous work
is associated with
assembled together. In that sense, some amount of re-use is at work.
states are largely generated through tinkering
have found them as the basic structures associated with specific
computations, have common motifs might support this view.
The fact that biological systems, also involved in performing
subsets. The final consequence is that the patterns of network
is actually a particularly relevant example. However, since
the statistical distribution of network motifs involves dealing
with large numbers of different subgraphs, the question of
how motifs in general might reflect functional traits requires
the formulation of appropriate null models of graph evolution.
Such models must ignore any functional trait in order to
test the possibility that the global properties of network
structure (such as graph heterogeneity) might strongly influence
what we should expect.

The model chosen here has been a duplication-rewiring
one [6]. These models have been shown to generate hetero-
genous graphs with many properties close to relevant biolog-
ical systems such as protein-protein interaction maps.
Network heterogeneity is largely due to effective preferential
attachment. Additionally, the rules of duplication strongly
bias the types of motifs to be formed towards some special
subsets. The final consequence is that the patterns of network
motifs generated by the duplication model might be able to explain
(statistical terms) the observed abundances of mo-
tifs, with no further requirement of functional constraints.
The fact that biological systems, also involved in performing
computations, have common motifs might support this view.
Although sharing common motifs seems to call for common
functionalities, it is important to remember that biological
structures are largely generated through tinkering [10,11].
Protein interaction networks grow by gene duplication, and
neural networks also experience increases of cell numbers
together with wide synaptic changes. Perhaps the common
traits are a by-product of the common tinkered evolution
based on extensive reuse and copy of available structures.

One final comment concerns with the common subgraphs
also shared by digital circuits. They are not obtained, strictly
speaking, through a process of duplication and rewiring. Al-
though the way complex circuits are built does include some
amount of reuse,² considerations involving low cost in links
are of fundamental importance. In spite of such constraints, it
has been shown that electronic circuits have small-world
structure and are also highly heterogeneous [30]. Previous
work seems to indicate that optimal design towards efficient

²As circuit complexity increases (both in terms of number of com-
ponents and computational tasks) it becomes more difficult to de-
sign from scratch choosing sets of small gates and building optimal,
low-cost circuits. Predefined gates involving well-known (and
sometimes complex) input-output functions are widely used and
assembled together. In that sense, some amount of re-use is at work.
communication at low cost can generate scale-free, heterogeneous architectures [31,32]. Such result suggests again that network heterogeneity might pervade motif abundances [33].

ACKNOWLEDGMENTS

We thank Maggie Fitzgerald, Frankie Dunn, and Eddie Scrap for useful input. We also thank Shalev Itzkovitz for a careful reading and comments on an earlier version of the manuscript. The analysis of network motifs has been done using available free software from Uri Alon’s Lab (seehttp://www.weizmann.ac.il/mcb/UriAlon/index.html). This work has been supported by Grant No. FIS2004-05422 and by the EU within the 6th Framework Programme under Contract No. 001907, “Dynamically Evolving, Large Scale Information Systems” (DELIS).

[20] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design Patterns Elements of Reusable Object-Oriented Software (Addison-Wesley, New York, 1994).
[25] All software maps analyzed here (and others studied by other authors) have been shown to be sparse. Correlations have been also analyzed in R. V. Solé and S. Valverde, in Complex Networks, edited by E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai, Lecture Notes in Physics (Springer, Berlin, 2004), pp. 169–190. Using statistical measures derived from information theory, it was shown that software maps are considerably uncorrelated.