Faster Detection of Network Motifs

Sebastian Wernicke

Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany

wernicke@minet.uni-jena.de

Definition of "Network Motif"

Motifs in a graph are small subgraphs which occur in significantly higher frequencies than in random networks with the same sequence of vertex degrees.

Motifs Are Useful For Uncovering Structural Design Principles...

The main idea is that "evolution preserves modules that define specific [...] functions" (Vespignani, 2003). The motif approach has led to interesting results in areas such as protein-protein interaction prediction and hierarchical network decomposition. As an example, motifs in the transcriptional networks of E. Coli and S. Cerevisiae can be attributed specific functionalities such as the generation of temporal expression programs.

...But Computationally Expensive To Find

Motif detection includes three subtasks: Finding subgraph classes, grouping them according to isomorphy, and determining their significance. The first task is expensive because the number of small subgraphs can be quite large even for small networks. The third task is expensive because (before this work) it involves the explicit generation of thousands of random graphs. (Efficient algorithms exist for the second subtask.)

We Propose A Faster Algorithm for Motif Detection

We propose a faster, unbiased algorithm for subgraph enumeration and -sampling. Also, we show how subgraph significance can be determined without the explicit generation of random graphs.

First Improvement: Fast and Unbiased Subgraph Sampling

The Previous Approach: In order to speed up network motif detection, Kashtan et al. propose a subgraph sampling algorithm. Their idea is that we start by selecting a random edge in the input graph and then randomly extend this subgraph until a connected subgraph with the desired number of vertices is obtained:

Algorithm: EDGE SAMPLING(G, k) (ESA)
Input: A graph $G = (V, E)$ and an integer $2 \leq k \leq |V|$.
Output: Vertices of a randomly chosen size-k subgraph in G.

```
01  {u, v} ← random edge from E
02  V' ← {u, v}
03  while |V'| ≠ k do
04      {u, v} ← random edge from V' × N(V')
05      V' ← V' ∪ {u} ∪ {v}
06  return V'
```

The problem is that this algorithm is biased. This requires an expensive corrective calculation and for some subgraphs the sampling quality is poor.

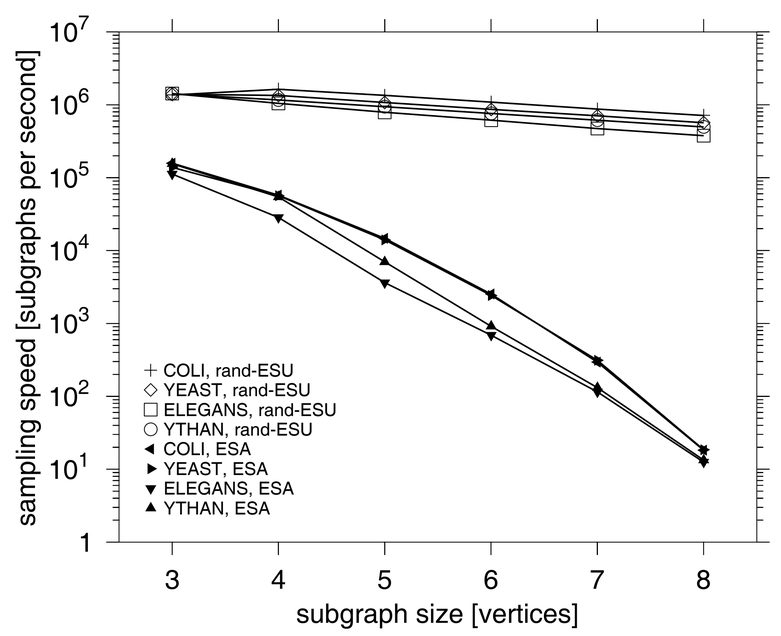
Our New Approach: We have developed a fast subgraph enumeration algorithm. This algorithm can be visualized in a tree-like structure. Random traversal of this "full enumeration tree" allows for efficient and unbiased subgraph sampling.

Algorithm: ENUMERATESUBGRAPHS(G, k) (ESU)
Input: A graph $G = (V, E)$ and an integer $1 \leq k \leq |V|$.
Output: All size-k subgraphs in G.

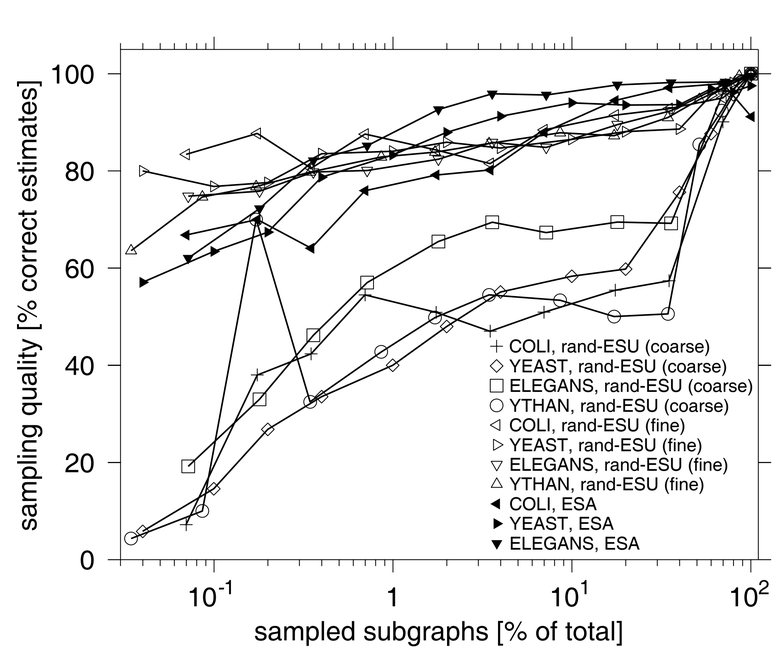
```
01  for each vertex v ∈ V do
02      V_Extension ← {u ∈ N(v) | u > v}
03      call EXTENDSUBGRAPH({v}, V_Extension)
04  endfor
```

EXTENDSUBGRAPH$(V_{Subgraph}, V_{Extension})$

```
E1  if |V_Subgraph| = k then output G[V_Subgraph] and return
E2  while V_Extension ≠ ∅ do
E3      Remove a vertex w from V_Extension
E4      V'_Extension ← V_Extension ∪ {u ∈ N_excl(w, V_Subgraph) | u > v}
E5      call EXTENDSUBGRAPH(V_Subgraph ∪ {w}, V'_Extension)
E6  return
```



Depending on subgraph size, our algorithm (called RAND-ESU) is some orders of magnitude faster than the fastest previous approach proposed by Kashtan et al. (called ESA).



The RAND-ESU algorithm yields a sampling quality that is comparable to ESA and more consistent for different graphs.

Second Improvement: Direct Calculation of Significance

In motif detection, the significance of a subgraph is determined by its mean concentration in random graphs with the same degree sequence. We have found a fast algorithm for determining subgraph significance without the need for an explicit generation of such random graphs. (Explicit generation was originally suggested by Milo et al.) Compared to previous approaches, we also gain the ability to analyze only specific subgraphs.

COLI $\langle C_k^i \rangle$	9.1e-1	3.7e-2	1.9e-4	5.0e-2	1.4e-3	2.1e-6	7.6e-8	3.4e-7	2.9e-6	2.9e-5	8.0e-7	–	–
$\langle \hat{C}_k^i \rangle$	9.0e-1	4.2e-2	2.6e-4	5.5e-2	1.4e-3	2.1e-6	1.3e-7	8.7e-8	2.3e-6	4.4e-5	1.1e-7	8e-12	6e-15
$\langle C_k^i \rangle / \langle \hat{C}_k^i \rangle$	1.0	0.9	0.7	0.9	1.0	1.0	0.6	3.9	1.3	0.7	7.4	–	–
YEAST $\langle C_k^i \rangle$	9.1e-1	3.7e-2	1.8e-4	5.0e-2	1.4e-3	9.5e-7	–	2.6e-7	2.3e-6	2.9e-5	3.4e-7	–	–
$\langle \hat{C}_k^i \rangle$	8.9e-1	3.0e-2	1.2e-4	7.6e-2	1.2e-3	1.5e-6	2.8e-8	4.4e-8	5.4e-7	1.0e-5	1.0e-7	1e-14	1e-15
$\langle C_k^i \rangle / \langle \hat{C}_k^i \rangle$	1.0	1.2	1.5	0.6	1.2	0.7	–	6.1	4.3	2.9	3.3	–	–
ELEG. $\langle C_k^i \rangle$	2.0e-1	3.3e-1	2.7e-2	3.7e-1	3.3e-2	1.7e-3	1.5e-3	2.0e-3	4.4e-3	2.9e-2	1.4e-3	3.8e-4	1.5e-5
$\langle \hat{C}_k^i \rangle$	2.0e-1	3.3e-1	2.9e-2	3.6e-1	3.6e-2	2.0e-3	1.9e-3	2.3e-3	4.7e-3	3.0e-2	1.5e-3	4.0e-4	1.5e-5
$\langle C_k^i \rangle / \langle \hat{C}_k^i \rangle$	1.0	1.0	0.9	1.0	0.9	0.9	0.8	0.9	0.9	1.0	0.9	0.9	1.0
YTHAN $\langle C_k^i \rangle$	4.1e-1	2.3e-1	3.3e-2	2.2e-1	5.1e-2	3.0e-3	2.7e-3	2.8e-3	2.0e-3	3.6e-2	5.3e-3	1.1e-3	5.8e-5
$\langle \hat{C}_k^i \rangle$	3.7e-1	2.4e-1	3.9e-2	2.2e-1	5.6e-2	3.5e-3	4.8e-3	5.0e-3	3.0e-3	5.2e-2	8.1e-3	2.7e-3	7.5e-4
$\langle C_k^i \rangle / \langle \hat{C}_k^i \rangle$	1.1	1.0	0.9	1.0	0.9	0.8	0.6	0.6	0.6	0.7	0.6	0.4	0.1

The mean subgraph concentrations in a random graph with given degree sequence (first lines) can be accurately estimated without the explicit generation of random graphs (second and third lines). This is much faster.

Friedrich-Schiller-Universität Jena
Institut für Informatik
Chair of Prof. Dr. Rolf Niedermeier

Supported by Deutsche Telekom Stiftung

Supported by National German Academic Foundation