Supplementary Information: Efficient Estimation of Graphlet Frequency Distributions in Protein-Protein Interaction Networks

Pržulj, N., Corneil, D. G., and Jurisica, I.[1]

December 3, 2005

[1]To whom correspondence should be addressed

Contents

We give here the supplementary information for the results presented in the paper. We first give an overview of our approaches. Then we briefly describe the hardware used to run the experiments. Finally, we describe the results of the two heuristic approaches.

1 Overview of Our Approaches

The principles of Time Limited Node Processing (TLNP) and Targeted Node Processing (TNP) heuristics are presented in Section 2.3.1 of the paper. The basics of the exhaustive search for 3- and 4-node graphlets is presented in Algorithm 1. 5-node graphlets were found in a similar way (the full description of how to detect them is long and thus is omitted). After we find graphlets in this way, each instance of P_3s, i.e., of graphlet 1 in Supplementary Figure 1, is counted twice – once from each end-node of the P_3. Thus, we divide the number of P_3s obtained in this way by 2. Similarly, each C_3 is counted 6 times, twice starting from each of its nodes and thus, the number of C_3s obtained in this way is divided by 6. We correct for over-counting of other graphlets in a similar way. Clearly, graphlets of type 29 in Supplementary Figure 1, i.e., K_5s, are the most severely over-counted with each one of them being counted $5! = 120$ times.

The basics of single node processing in TLNP approach to detect 3- and 4-node graphlets containing one particular node are presented in Algorithm 2. The 5-node graphlets are detected using similar principles (the full description of how to detect them is long and thus is omitted). As in the exhaustive search (Algorithm 1), we correct for overcounting of graphlets (see above). Note that if we only process a small percentage of network nodes, these corrections may not be quite accurate. However, they still provided good approximate solutions.

For the TNP heuristic, we use a stable sort to sort the nodes first by increasing degree and then by decreasing eccentricity (eccentricity is defined in Section 2.1 of the paper). Examples showing degrees and eccentricities of the top 2% of nodes sorted in this way are presented in Supplementary Table 1.

Algorithm 3 describes the NLS approach. This approach is based on Mathon's Restricted Neighborhood Local Search (Mathon 2004). The description of the results and their dependence on the choice of parameters is presented in Section 3.2.

Algorithm 1: FIND GRAPHLETS EXHAUSTIVELY(G)

```
for all nodes u of G do
    for all adjacent nodes v of node u do
        for all adjacent nodes w of v such that w ≠ u do
            (*finding 3-node graphlets*);
            H ← subgraph of G induced on {u, v, w};
            if H has 2 edges then
                H is a path P3, i.e., an instance of graphlet 1
            end
            if H has 3 edges then
                H is a cycle C3, i.e., an instance of graphlet 2
            end
            for all adjacent nodes x of w such that x ≠ u and x ≠ v do
                (*finding 4-node graphlets*);
                H ← subgraph of G induced on {u, v, w, x};
                if H has 6 edges then
                    H is a K4, i.e., an instance of graphlet 8
                end
                if H has 5 edges then
                    H is an instance of graphlet 7
                end
                if H has 4 edges then
                    if H has a degree 3 node then
                        H is an instance of graphlet 6
                    else
                        H is a 4-cycle, i.e., an instance of graphlet 5
                    end
                end
                if H has 3 edges then
                    H is a path, i.e., an instance of graphlet 3
                end
            end
        end
        (*detect Claws, i.e., graphlets of type 4*);
        for all nodes v, w, x adjacent to u which are all different do
            H ← subgraph of G induced on {u, v, w, x};
            if H has 3 edges then
                H is a Claw, i.e., an instance of graphlet 4
            end
        end
    end
end
```

Algorithm 2: TLNP(G,u)

```
for all adjacent nodes v of node u do
    for all adjacent nodes w of v such that w ≠ u do
        (*finding 3-node graphlets*);
        H ← subgraph of G induced on {u, v, w};
        if H has 2 edges then
            H is a path P3, i.e., an instance of graphlet 1
        end
        if H has 3 edges then
            H is a cycle C3, i.e., an instance of graphlet 2
        end
        for all adjacent nodes x of w such that x ≠ u and x ≠ v do
            (*finding 4-node graphlets*);
            H ← subgraph of G induced on {u, v, w, x};
            if H has 6 edges then
                H is a K4, i.e., an instance of graphlet 8
            end
            if H has 5 edges then
                H is an instance of graphlet 7
            end
            if H has 4 edges then
                if H has a degree 3 node then
                    H is an instance of graphlet 6
                else
                    H is a 4-cycle, i.e., an instance of graphlet 5
                end
            end
            if H has 3 edges then
                H is a path, i.e., an instance of graphlet 3
            end
        end
    end
    (*detect Claws, i.e., graphlets of type 4*);
    for all nodes v, w, x adjacent to u which are all different do
        H ← subgraph of G induced on {u, v, w, x};
        if H has 3 edges then
            H is a Claw, i.e., an instance of graphlet 4
        end
    end
end
```

Algorithm 3: NLS(G,n,m,NUM-EXP,NUM-MOVES,DIV-FREQ,DIV-DUR)

```
for k ∈ {1,...,NUM-EXP} do
    Pick a node v of G at random;
    for i ∈ {1,...,n−1} do
        Neighbors ← {all nodes in the i^th neighborhood of v} ∪ {v}
    end
    Subnodes ← n random connected nodes from Neighbors;
    Form a subgraph G_s of G induced on the nodes from Subnodes;
    if G_s has m edges then
        return G_s
    else
        (*Begin moves*);
        for j ∈ {1,...,NUM-MOVES} do
            if j mod DIV-FREQ ≥ DIV-DUR then
                (*Take a locally optimal move*);
                Pick a random node u from G_s;
                Pick a random node v from the neighborhood of G_s;
                G_s ← G_s \ {u} ∪ {v} (*i.e., swap u and v*);
                if number of edges of G_s is closer to m then
                    keep this G_s;
                    if G_s has m edges then
                        Return G_s
                    end
                else
                    restore old G_s;
                    forbid node u from being picked from G_s in the next
                    move;
                end
            else
                (*Diversify*);
                Pick a random node u from G_s;
                Pick a random node v from the neighborhood of G_s;
                G_s ← G_s \ {u} ∪ {v} (*i.e., swap u and v*);
            end
        end
    end
end
Return G_s
```

2 Hardware Benchmarks

We ran the exhaustive graphlet search algorithm on the following two multi-processor University of Toronto Computer Science Laboratory (CSLab) server machines:

- `parse.cs Asus A7M-266D (2xAthlonMP1900+) RedHat Linux 7.3 (2.4.x kernel);`
- `ponder.cs Dell PowerEdge 2650 (2xP4Xeon-2.80GHz) RedHat Linux 7.3 (2.4.x kernel).`

Our graphlet finding algorithm runs 49% faster on `parse.cs` than on `ponder.cs`. However, since these are general-purpose compute servers, we ran our experiments on a server which was less heavily used at the time (which was mostly `ponder.cs`). We used C++ programming language, LEDA C++ library (Mehlhorn and Naher 1999), and version 2.9.2 of g++ compiler for implementing all of our algorithms. Note that the running times taken by our programs could by significantly reduced by using a newer version of the g++ compiler; however, since we are interested in speed-up factors of our heuristics rather than in the absolute values of the search times, we did not re-compile our programs with a newer compiler and re-run our experiments.

We ran the heuristic graphlet search experiments either on the `ponder` server, or on the University of Toronto Computing Disciplines Facility (CDF) distributed machine. The CDF distributed machine is a cluster of Linux machines out of which several are usually down and thus the heuristic graphlet search experiments were run on the machines which were available at the time. This resulted in negligible variation in run times which can safely be ignored. Compared with the same heuristic graphlet search algorithm run on `ponder.cs`, CDF distributed machine is $3-6\%$ slower. Thus, for our purposes, we consider the running time of the CDF distributed machine to be roughly the same as the running time of the `ponder.cs` server.

For comparing running time of exhaustive and heuristic algorithms, we scaled all running times to correspond to `ponder.cs`. The loading and output times are included in the total reported processing times.

3 Results

3.1 Time Limited Node Processing (TLNP) and Targeted Node Processing (TNP)

We first describe the results of this approach with limited processing times placed on each node (TLNP). Then we give statistics on the degree and eccentricity properties of the finished and unfinished nodes. We do a full, time unlimited processing

of selected nodes next (TNP) and report the results. The node processing time limits that we use in the TLNP approach are 3 seconds, 10 seconds, 30 seconds, and 3 minutes. We present the results of the TLNP and TNP approaches for PPI and the corresponding geometric random networks, as well as the results of the TLNP for the corresponding SF and ER-DD networks. We did not do TNP experiments for SF and ER-DD networks, since the results of the TLNP experiments indicated that such an approach would not work well for these networks. In the end, we present the results of applying this approach to very large geometric random networks.

3.1.1 PPI Networks

We describe the results of the Time-Limited Node Processing (TLNP) heuristic for the four PPI networks: the yeast *S. cerevisiae* high confidence and the top 11000 PPI networks (von Mering et al. 2002), and the fruitfly *D. melanogaster* high confidence and the entire PPI networks (Giot et al. 2003). These networks contain 2455 interactions (edges) between 988 proteins (nodes), 11000 interactions between 2401 proteins, 4637 interactions between 4602 proteins, and 20007 interactions between 6985 proteins, respectively. We describe general patterns in graphlet estimates produced by this heuristic applied to PPI networks.

The numbers and percentages of unfinished nodes for each of these four data networks and for different time bounds of the TLNP algorithm are presented in Supplementary Table 2. Percentages vary from 2.6% to over 56%. Also, the relative graphlet frequency distances between the estimated and the exact graphlet counts are shown (distance is defined in Section 2.4 of the paper). Graphlet frequency distributions of the heuristic and exact searches are presented in Supplementary Figures 3 and 4 (also in Supplementary Figure 2). These figures indicate that most of the graphlets are consistently under-counted by this heuristic approach. To see this more clearly, we computed fractions of graphlets of type i found by the heuristic algorithm with respect to the exact number of graphlets of type i, i.e., we computed $\frac{H_i(G)}{N_i(G)}$, $1 \leq i \leq 29$, where $H_i(G)$ is the number of graphlets of type i in G found by the heuristic search algorithm, and $N_i(G)$ is the exact number of graphlets of type i in G ($i \in \{1, \ldots, 29\}$). These fractions for yeast and fly PPI networks are presented in Supplementary Tables 3 and 4. (The graphlet numbering scheme is as in Supplementary Figure 1.)

As Supplementary Figures 3-4 and Supplementary Tables 3-4 illustrate, most of the graphlets in these PPI networks get uniformly under-counted. The graphlets that are the most severely under-counted in each of these networks are graphlets 4, 10, 11, and 14. As mentioned in the paper, all of these graphlets contain graphlet 4 as an induced subgraph, which is expected, since, high degree nodes, as well as nodes in dense neighborhoods, get under-counted by this heuristic. Thus, we do

not expect that this heuristic graphlet search approach would work well on network models with pronounced hub nodes, such as scale-free networks. However, despite the presence of hubs in PPI networks, it works surprisingly well for these networks. This further supports our hypothesis that PPI networks have a different local structure than scale-free networks. In some cases, larger graphlets that contain some of these under-counted graphlets as subgraphs are also heavily under-counted. For example, in the larger yeast PPI network, under-counting of graphlets 5, 7, 8, 20, 22, and 24-29 also happens, all of which, except for graphlet 5, contain graphlet 4 as a subgraph.

It is interesting to notice that for all four PPI networks, $10 - 15\%$ of unprocessed nodes consistently yield relative graphlet distances from the exact counts of 33-39. These low distances (the definition of "low distance" is given in the paper) indicate that this heuristic approach gives a reasonably good approximation of the distribution of graphlet frequencies, despite the under-counting (also can be seen in Supplementary Figures 3 and 4). That is, the graphlet distribution that TLNP produces for PPI networks is a good approximation of the exact graphlet distribution except for a translation, i.e., almost uniform under-counting of graphlets.

We examined the properties of the nodes that did not get processed by these TLNP experiments. In addition to finding the degrees of these nodes, we wanted to find out their location in the network. The radii, diameters and degree statistics of the five PPI networks are presented in Supplementary Table 5. Since these networks are disconnected and consist of many small connected components with the largest connected component containing most of the nodes and edges, we report the radii and diameters of the largest connected components in these networks. All nodes unfinished by the TLNP experiments with all tested cut-off times belong to the largest connected component of the corresponding PPI network. The eccentricity and degree statistics of the nodes unfinished by the TLNP experiments with different time cut-offs are presented in Supplementary Table 6. Note that in each of the PPI networks and in each of the TLNP experiments, the eccentricities of unfinished nodes are close to the radius of the largest connected component of the corresponding PPI network (Supplementary Tables 5 and 6). Thus, unfinished nodes are close to the center of the PPI network. Also, the average degree of unfinished nodes in each of the experiments is much higher than the average degree of the PPI network (Supplementary Tables 5 and 6). Therefore, as expected, most of the nodes that remain unprocessed by TLNP in PPI networks are of high degree and deep inside, i.e., close to the center of the network.

Since nodes that remain unprocessed by the TLNP are mainly of high degree and close to the center of the network, we proposed the TNP strategy to estimate the pattern of graphlet distribution in PPI networks as described in the paper. We tested the TNP approach on the yeast high-confidence PPI network. We processed

the top 10%, 20%, 30%, 40%, and 50% of the nodes of the yeast high-confidence PPI network ordered by a stable sort first in increasing degree order and then in decreasing eccentricity order. That is, we did not initiate a search at 90%, 80%, 70%, 60%, and 50% of the nodes in this PPI network, respectively. The time to process the selected nodes was 6 seconds, 47 seconds, 3 minutes and 40 seconds, 11 minutes and 13 seconds, and 66 minutes and 19 seconds, respectively (the exhaustive search takes almost 9 hours), resulting in graphlet distribution distances from the exact counts of 68.34, 45.91, 39.37, 39.80, and 16.59, respectively (Supplementary Figure 2 E and Supplementary Table 7). The distance of 45.91 obtained by processing 20% of the sorted nodes is small and not much is gained in accuracy until we get to process 50% of the nodes. The ratio of exhaustive and this heuristic search time is $r = \frac{T_E}{T_H} = \frac{8h57m16.63s}{46.72s} = \frac{32236.63}{46.72} \approx 690$. A description of how well the processing of the nodes that are on the fringes of the model networks approximates the graphlet frequency distribution of model networks is presented in Section 3.1.2.

3.1.2 Results for Model Networks

We are interested in determining how well the TLNP heuristic graphlet search approximates graphlet frequency distributions of model networks. In particular, since we believe that PPI networks are best modeled by geometric random graphs (Pržulj, Corneil, and Jurisica 2004; Pržulj 2005), we are interested in the performance of the TLNP on geometric random graphs. First, we examine its performance on geometric random graphs with sizes and densities comparable to the sizes and densities of PPI networks. Then, we examine its performance on a very large geometric random graph (Section 3.1.4).

We also briefly discuss the performance of this algorithm on ER-DD and SF networks (defined in the paper). Since these models do not approximate well local structural properties of PPI networks, we do not focus on finding a heuristic approach that works well for approximating the graphlet distribution of these model networks.

Geometric Random Graphs We first applied the TLNP heuristic with various cut-off times to geometric random graphs corresponding to the two yeast PPI networks. The numbers of unfinished nodes and the distances between the estimated and the exact graphlet counts are presented in Supplementary Tables 8 and 9. 2-dimensional (GEO-2D), 3-dimensional (GEO-3D), and 4-dimensional (GEO-4D) geometric random graphs corresponding to the yeast high confidence PPI network are easily fully, or almost fully processed by this heuristic with tested cut-off times (Supplementary Table 8). In geometric networks with higher densities of edges

than in the yeast high confidence PPI network, a significant percentage of nodes remains unprocessed (Supplementary Table 8). It is interesting to notice that for these networks, even when a very large percentage of nodes remains unprocessed, the resulting graphlet frequency distribution pattern is very close to the exact one (Supplementary Figures 5 - 9). For example, the TLNP with 3 second node processing time cut-off applied to the three GEO-3D networks with the same number of nodes, but three times as many edges as the PPI network resulted in $55.36-66.90\%$ of unprocessed nodes; despite this large percentage of unprocessed nodes, the distances from the exact graphlet counts were only between 8.29 and 9.68 (Supplementary Table 8 and Supplementary Figure 5). This effect is even more dramatic for the networks with six times as many edges as the PPI network, since for these networks, due to their larger densities, the chosen cut-off times produced larger numbers of unprocessed nodes; even when over 98% of the nodes in these networks remain unprocessed, the distance between the heuristic and exact graphlet distributions is only between 32.76 and 67.22 (Supplementary Table 8 and Supplementary Figure 9: 3 second time limited TLNP for GEO-3D-6x graphs). Note that the distance of 67.22 happens when we process only 2 out of 988 nodes! The remaining four distances were between 32.76 and 40.60 and they resulted from processing only between 3 and 11 out of 988 nodes of a network. If we process only between $11.91-20.24\%$ of nodes in these networks, the distance falls to $14.80-17.84$ (Supplementary Table 8 and Supplementary Figure 9: 30 second time limited TLNP for GEO-3D-6x graphs). Similarly, not processing $12.63 - 20.98\%$ of nodes of 2-, 3-, and 4-dimensional geometric random networks corresponding to the yeast top 11000 PPI network results in tiny graphlet distribution distances of $3.60 - 6.05$ between the heuristic and the exact graphlet counts (Supplementary Table 9 and Supplementary Figures 6 - 8).

As stated in the paper, the following pattern emerges from this analysis. The under-counting of graphlets in geometric random networks is very uniform and, thus, it is enough to process only a small fraction of nodes to get good estimates of the graphlet frequency distribution patterns for these networks, apart from translation. That is, the graphlet frequency distribution obtained in this heuristic way multiplied by almost a constant should give a good estimate of the exact graphlet frequency distribution in these networks. For example, if we multiply each of the 3 second time bounded TLNP estimated graphlet frequency distributions of the three GEO-3D networks (presented in Supplementary Figure 5) by 7, the distances between the exact and the corrected estimates of the graphlet distribution stay the same, and the resulting corrected estimated graphlet frequencies are in good alignment with the exact ones (Supplementary Figure 10).

The question again is, which nodes to process, i.e., can we do a quick, targeted node processing to get a good estimate of the graphlet distribution? As before, to

answer this question, we examined the degrees and eccentricities of processed (finished) and unprocessed (unfinished) nodes in these geometric random networks. Radii, diameters and degree statistics of GEO-3D-3x and GEO-3D-6x networks corresponding to the yeast high-confidence PPI network are presented in Supplementary Table 10. Note that all of these networks are connected. Eccentricity and degree statistics of processed and of unfinished nodes (by the TLNP with different time cut-offs) in GEO-3D-3x networks corresponding to the yeast high confidence PPI network are presented in Supplementary Tables 11 and 12, respectively. Similarly, eccentricity and degree statistics of processed and of unfinished nodes by the TLNP with different time cut-offs in GEO-3D-6x networks corresponding to the yeast high confidence PPI network are presented in Supplementary Tables 13 and 14, respectively. In each of the experiments, the average degree of unprocessed nodes is higher than the average degree of processed nodes, while the average eccentricity of unprocessed nodes is lower than the average eccentricity of processed nodes. Thus, the nodes that do not get processed in each of these networks and with each of the tested TLNP cut-off times are deeper in the network and of higher degree than the nodes that get processed. As we increase the processing cut-off time and allow more and more nodes to get processed, the average degree of both processed nodes and unprocessed nodes grows, while the average eccentricities fall (Supplementary Tables 11 - 14). This means that with increasing cut-off times, we are able to process nodes of higher degree that are "deeper" in the network. The same holds for geometric random networks corresponding to the yeast top 11000 PPI network (Supplementary Tables 15 and 16).

As mentioned in the paper, a possible explanation of why this heuristic approach works so well on geometric random networks is the following. In this heuristic, we are starting from the nodes on the fringe of the network and "grabbing" a sample of graphlets that are up to depth 5 from the fringe of the network. Since the structure of these networks is uniform inside the network (note that the boundary has a different structure), it seems to be enough to sample the graphlets that are about 5-deep from the fringe of the network to get the estimate of the distribution of graphlets in the whole network. It can be argued that these networks are of small diameter, so by going "5-deep" into the network, we may be reaching the center of the network. However, sampling the center may not even be needed, since the structure of these networks looks the same in all inner parts of the network. This is further supported by the observation that this approach approximates well the graphlet distributions of geometric networks with diameters of 52-53 (Supplementary Tables 9 and 15 and Supplementary Figure 6).

These observations lead to the same strategy for estimating graphlet frequency distributions in geometric random networks as before, for PPI networks. We need to select a fraction of nodes of a geometric random network that are on the network

"boundary" and are of low degree and to find graphlets in the neighborhood of these nodes only. The fraction of nodes that can be processed in a specified time will depend on the size and density of the network: the denser the network, the higher is its average degree and thus we can process fewer of its nodes in a specified amount of time.

We tested this TNP approach on the five 3-dimensional geometric random networks corresponding to the yeast high-confidence PPI network with 6 times as many edges as the PPI network. As before, we used a stable sort to sort the nodes of these networks, first in increasing degree order and then in decreasing eccentricity order (an example of node sorting is presented in Supplementary Table 1). We selected the top 1% of such ordered nodes of these networks and fully processed them. The resulting heuristic graphlet frequency distributions are in very good agreement with the results of the exhaustive search (see Supplementary Figure 11 and Supplementary Tables 17 and 18). Note that it took 2-4 minutes to process the nine selected nodes in each of these networks and the resulting graphlet frequency distributions are at distance of 32-51 from the result of the exhaustive search. This is a big time saving when compared to the total CPU time of around 49 minutes to perform the 3 second time-limited TLNP, and even bigger time saving when compared to the processing time of 18.5-21.5 hours of the exhaustive searches, without a decrease in the quality of the estimated graphlet frequency distribution patterns (complete run time results are presented in Section 3.1.3). For example, the total CPU time for TLNP with 1% of targeted nodes for the GEO-3D-6x 1 network was 3 minutes and 46.75 seconds, while the exhaustive search on this network took 18 hours and 52 minutes. Thus, for this network, $r = \frac{T_E}{T_H} = \frac{18h52m}{3m46.75s} = \frac{67920}{226.75} \approx 300$. When we choose to process the top 2% of the nodes sorted as above, we obtain more uniform distances from the exact counts (35-41) at the expense of almost tripling the processing time and $r \approx 151$ when compared with processing time of the top 1% of the sorted nodes (Supplementary Table 18).

As previously mentioned, more experimentation with a larger number of networks is needed to determine better node selection criteria that would further decrease the processing time and possibly increase the quality of the estimated graphlet distributions. Also, the dependence of graph density, node selection, and processing time needs to be understood. The "translation" of the estimated graphlet distribution and its "alignment" with the exact one is needed as well. The results so far suggest that these translations and their dependence on the fraction of processed nodes and network type and density should be easy to find.

Erdös-Rényi (ER-DD) and Scale-Free (SF) Graphs Under-counting of graphlets in the SF and ER-DD model networks (described in the paper) is not uniform and

thus it results in higher graphlet distances between the exact and the estimated graphlet counts, despite the small number of unprocessed nodes. To demonstrate this, we performed TLNP experiments with 3 minute node processing cut-off time on ER-DD and SF networks corresponding to the four PPI networks that we analyzed as well as the worm *C. elegans* PPI network (Li, Armstrong, Bertin, Ge, Milstein, Boxem, Vidalain, Han, Chesneau, Hao, Goldberg, Martinez, Rual, Lamesch, Xu, Tewari, Wong, Zhang, Berriz, Jacotot, Vaglio, Reboul, Hirozane-Kishikawa, Li, Gabel, Elewa, Baumgartner, Rose, Yu, Bosak, Sequerra, Fraser, Mango, Saxton, Strome, van den Heuvel, Piano, Vandenhaute, Sardet, Gerstein, Doucette-Stamm, Gunsalus, Harper, Cusick, Roth, Hill, and Vidal 2004) (the description of how we constructed these networks is given in (Pržulj, Corneil, and Jurisica 2004)). Erdös-Rényi (ER) networks corresponding to these PPI networks got fully processed by this heuristic (note that these networks have a Poisson degree distribution and lack hubs, i.e., they exhibit "uniformity" of structure), and thus we excluded them from further analysis.

A total of nine SF and thirteen ER-DD networks were analyzed. Due to the common properties captured by a network model, all networks of the same model with similar edge densities will behave very similarly and such behaviors can be analytically proven; this is why it is sufficient to exhibit a behavior of a certain network model on very few examples. The statistics for the number and fraction of unprocessed nodes and the distances between the exact and the heuristic graphlet search results are presented in Supplementary Table 19. Supplementary Figures 12 - 13 present graphlet frequency distributions obtained by the exact and the heuristic graphlet searches for SF networks while Supplementary Figures 14 - 16 present those for ER-DD networks. Despite a tiny percentage of nodes remaining unprocessed in these networks, the graphlet frequency distances between the exact and the heuristic results are large. This is due to severe under-counting of graphlets 4, 10, 11, and 14 (Supplementary Figures 12 - 16).

Note that a different distance measure which would depict better the variation between the estimated counts for different graphlets would provide even better support for these observations. However, since Supplementary Figures 12 - 16 provide a good illustration of this effect, we leave the design of such a distance measure for future research.

Similar undercounting of graphlets 4, 10, 11, and 14 was observed in PPI networks, but not to such a large extent, despite the presence of hubs in PPI networks. Thus, this provides further support to our previous observation that despite the presence of hubs in PPI networks, their local structure is closer to the local structure of GEO networks than to the local structure of SF networks.

As before, statistics on degrees and eccentricities of unprocessed nodes, when compared to the diameters and radii of the networks they belong to, indicate that

the nodes that remained unprocessed are mainly of high degree and deep inside the network (Supplementary Tables 20 and 21). (Note that degree statistics of ER-DD nets in Supplementary Table 20 are all the same, since these networks were constructed to have exactly the same degree distribution as the corresponding PPI networks.)

3.1.3 Run Times

In practice, the running times of TLNP experiments are small compared to the running times of the exhaustive searches. For example, the times taken by the exhaustive and the TLNP heuristic searches with different cut-off times for the five PPI networks are presented in Supplementary Table 22. The heuristic times reported in the table are the sums of the CPU processing times taken by the nodes which were fully processed (i.e., finished) by these heuristic experiments. We intentionally excluded the times taken by the unprocessed nodes, because the idea of this approach is to do full processing of a percentage of targeted nodes; as seen before, low-degree peripheral nodes of the network are more easily processed than the high-degree central nodes and, thus, the low-degree peripheral nodes are to be chosen for processing by this heuristic (Supplementary Figure 11, Supplementary Figure 2 E, and Supplementary Tables 18 and 7). However, since due to the lack of time and CPU cycles we did not do many TNP experiments, we report here the run times of the nodes processed by the TLNP experiments with specified node processing time cut-offs. These run times should be comparable with the run times of the TNP experiments (see Supplementary Table 17 for an example).

As expected, by decreasing the TLNP cut-off time, less processing time is needed, fewer nodes get processed and the accuracy of the graphlet counts falls (Supplementary Table 22). The only exception is the fly high-confidence (FH) PPI network; since the percentage of unprocessed nodes for this network is extremely small, and since distributing the processes over multiple CPUs results in an overhead, the time taken by the heuristic is slightly larger than the time taken by the exhaustive search. Also note that the FH PPI network is very sparse and has no prominent hubs, and thus, it is easily processed by the exhaustive search. Therefore, clearly, this approach should be used for networks which are hard to process exhaustively and for which a larger percentage of nodes can be discarded from processing without a large penalty in accuracy.

This brings us to the geometric random graph model of PPI networks: as seen in Section 3.1.2, the accuracy of the TLNP estimated graphlet frequency distribution pattern for geometric random graphs is not greatly affected even when we do not process as many as 88%, or even 98%, of the nodes (Supplementary Table 8 and Supplementary Figure 9). Thus, a nice thing about this approach is that, for geo-

metric random graphs, and also for high confidence PPI networks (which, we have seen, are modeled well by geometric random graphs (Pržulj, Corneil, and Jurisica 2004)), we can choose almost an arbitrarily small percentage of their peripheral, low-degree nodes to process without adversely affecting the accuracy of the resulting estimated graphlet frequency distribution, apart from translation, which should be easy to correct for. An example of the processing times taken by the exhaustive and the TLNP search algorithm with different cut-off times, percentages of unprocessed nodes, and distances between the exhaustive and the TLNP heuristic graphlet counts for geometric random graphs are presented in Supplementary Table 23; the time of 18-22 hours taken but the exhaustive graphlet search in these networks can be reduced to 13-21 seconds by the TLNP heuristic search with the resulting graphlet distribution distance of only 35-41.

As mentioned before, TLNP does not work well for estimating the graphlet distributions of scale-free networks. A sample of the processing times taken by the exhaustive search and the TLNP experiments with 3 minute cut-off time is presented in Supplementary Table 24. Clearly, a small number of hub nodes takes a vast majority of the graphlet search time in these networks, which results in severe under-counting in TLNP of "hub-specific" graphlets, i.e., graphlets with induced graphlet 4 as subgraphs (see Supplementary Tables 20 - 21 and 24, and Supplementary Figures 12 - 13). For example, the 5 hub nodes in the center of the network SF 5 (see Supplementary Tables 20, 21, and 24) take more than 6 hours to process; in contract, the remaining 99.5% of the nodes of the network take only 26 minutes and 24 seconds to process (Supplementary Table 24). Note that the resulting graphlet frequency distribution distances are over 20, which is quite large considering that well below 1% of nodes in these networks remain unfinished. A similar situation occurs for the ER-DD networks (see Section 3.1.2 and Supplementary Table 25). Note that ER networks are easily fully processed due to the lack of hubs (for example, each of the five ER networks corresponding to the yeast high-confidence PPI network are exhaustively processed in less than 3 minutes).

3.1.4 Discussion and Application to Very Large Geometric Random Networks

We have observed that the most frequently under-counted graphlets in this heuristic approach applied on PPI, ER-DD, and SF model networks are graphlets 4, 10, 11, and 14 (Supplementary Figure 1); for geometric random graphs the under-counting is uniform over all graphlets. Graphlets 4, 10, 11, and 14 all have graphlet 4 (also called a "Claw", or a "3-star") as an induced subgraph. Since this heuristic approach discriminates against counting graphlets in the vicinity of hubs, i.e., nodes of high degree, the graphlets with star-shaped underlying structure get under-

counted. The network models with the largest number of "stars" are the SF and ER-DD models (hubs have a large number of induced star-shaped subgraphs in their neighborhoods), and this is why we see severe under-counting of the graphlets containing stars when compared to under-counting of other graphlets in these models. This effect is seen in PPI networks, but not to such a large extent as in the SF and ER-DD networks. PPI networks behave more like the corresponding geometric random graphs, with a deviation exhibited in slight under-counting of graphlets with induced Claws; the under-counting of other graphlets is fairly uniform in these networks. Note that even the subgraphs 7, 20, 22, 25-29, that were slightly under-counted in the larger yeast PPI network, all have Claws as subgraphs; however, most of them also contain many triangles, again suggesting that denser regions of the network are hard to process using this heuristic.

To conclude, we have observed that this heuristic approach for estimating the number of graphlets in a network works well for geometric random graphs and not well for network models with hubs. However, it works surprisingly well for PPI networks despite the fact that they have hubs. Thus, if our hypothesis that the true structure of PPI networks, once we obtain more complete data on them, is similar to the structure of geometric random graphs, this heuristic approach will be adequate for estimating the graphlet distribution pattern in PPI networks and will result in uniform underestimation of the number of graphlets. In addition, with a decreased fraction of nodes that get processed and thus decreased processing time as well, the accuracy of the graphlet distribution estimate hardly decreases, which makes this approach appealing.

As discussed previously (see the paper), PPI networks for higher organisms will be much larger, possibly including hundreds of thousands of nodes and edges. This volume of data will be impossible to process exhaustively to find their graphlet distributions. Thus, heuristics will have to be used to estimate graphlet distributions in these networks.

To evaluate the scaling of the TLNP heuristic approach (which implies scaling of the TNP approach) to large networks, we tested it on a 3-dimensional geometric random graph with 100,000 nodes and 750,000 edges. This networks has 3 times as many edges as the two yeast PPI networks that we analyzed. We obtained two samples of processed nodes by testing two different cut-off times, 60-second and 90-second (Supplementary Table 26 and Supplementary Figure 17). These two cut-off times yielded 1.50% and 4.29% of finished nodes, respectively. The total processing times for the finished nodes were 16h 44m 23.27s and 2d 16h 24m 28.34s for the 60-second and 90-second cut-off times, respectively. Since it is not possible to obtain the exact graphlet distribution of this large geometric network, we compared the graphlet frequency distributions resulting from these two experiments with the exact graphlet frequency distributions of GEO-3D networks cor-

responding to the yeast high-confidence PPI network with 3 and 6 times as many edges as the PPI network (Supplementary Figure 17 and Supplementary Table 26). The distances between these exact and the estimated large network graphlet distributions are very low (23-29 from the 3 times denser, and 26-31 from the 6 times denser GEO-3D network corresponding to the yeast high-confidence PPI network) despite a tiny percentage of nodes being processed. As before, the finished nodes are of low degree and on the boundary of the network.

Similarly, we evaluated the scaling of the TLNP heuristic approach on perturbed GEO-3D networks to account for the noise present in the data. The same large GEO-3D network as above was used, but with 10%, 20%, and 30% of its edges being rewired at random, respectively. We obtained three different networks with 10% of rewired edges, by doing independent rewiring experiments three times. We did the same with 20% and 30% of the edges being rewired and obtained six more rewired networks. The results of running TLNP on these networks are presented in Supplementary Table 26 and Supplementary Figure 18. By increasing the level of noise in the geometric network, more processing time was needed to process the same percentage of nodes, and the distances of the estimated graphlet distributions from the exact graphlet distributions of a geometric random network grew. The estimated graphlet frequency distributions in these networks are becoming more ER-DD-like as the level of noise grows (Supplementary Figure 18).

These results indicate that the TLNP heuristic (and subsequently the TNP heuristic) can be used to determine the general pattern of graphlet distributions of large networks with geometric random graph structure. More experimentation with such large networks of different edge densities is needed to determine how the resulting estimated graphlet distributions should be scaled to give a good approximation of the exact graphlet distributions. The analysis described above demonstrates that the TLNP approach scales to large geometric random graphs.

3.2 Neighborhood Local Search (NLS)

We analyzed the yeast high-confidence PPI network and the corresponding model networks using the NLS approach (Algorithm 3 in Section 1) with the following choice of search parameters: maximum number of experiments is 2, maximum number of moves per experiment is 5, diversification frequency is 3, and diversification duration is 1 (i.e., every third move is random in the neighborhood of a selected nodes). We experimented with different numbers of seed nodes: for each graph $G(V, E)$ processed by this heuristic, we performed experiments using $|V|/8$, $|V|/4$, $|V|/2$, $|V|$ and $2|V|$ seed nodes per n-node, m-edge graphlet, respectively. We performed 10 distinct runs of the algorithm for each choice of the number of

seed nodes for each graph; that is, for every graph G, we ran the algorithm 10 times for each of the 5 above mentioned seed node numbers resulting in $10 \times 5 = 50$ different runs. The averages and standard deviations of estimated graphlet frequencies were obtained for the 10 runs for the same graph and the same number of seed nodes; the standard deviations were several orders of magnitude smaller than the corresponding averages. The resulting pattern of averages of graphlet frequency distribution estimates for the PPI network and the corresponding geometric model networks is close to the pattern of the exact graphlet frequency distributions for these networks (see Supplementary Figure 21 A and Supplementary Table 27 below). However, this is not the case for the ER, ER-DD, and SF model networks (Supplementary Figure 21 E and F, and Supplementary Table 27). Determining how to analytically bring the estimated graphlet frequencies to the exact ones in absolute values is left for future research.

The reason why this heuristic approach works much better for the PPI and GEO networks than for the ER, ER-DD, and SF networks lies in the following facts. In PPI and GEO networks, the frequencies of different graphlets are much more evenly distributed than in the ER, ER-DD, and SF networks. That is, in ER, ER-DD, and SF networks, the number of sparse graphlets is several orders of magnitude larger than the number of dense graphlets. This is not the case in PPI and GEO networks. Thus, since the algorithm is always trying to sample the same number of n-node, m-edge graphlets, the disproportionality of graphlet counts in ER, ER-DD, and SF networks cannot be fully detected by this heuristic algorithm.

Determining how many samples are enough is well explored in random sampling from databases and population statistics (see the paper). Note that we obtained very good graphlet frequency distribution estimates with as few as $\frac{|V|}{8}$ samples per n-node, m-edge graphlet on the yeast high-confidence PPI and the corresponding model networks with around 1000 nodes and 2400 edges. This is orders of magnitude smaller number of samples than the 10^5 samples that were required by the Kashtan et al. (2004) algorithm for much smaller *E. coli* transcriptional and *C. elegans* neural networks. We are doing a limited, 5-move search in the neighborhood of a random graphlet rather than just selecting a random graphlet; Kashtan et al. (2004) are correcting for non-uniform sampling by calculating probabilities to sample a random graphlet instead. Also, we are estimating only the graphlet frequencies relative to one another for the time being; an analytical "translation" of the resulting estimate should be easy to determine experimentally.

In order to determine the dependence of the obtained approximate graphlet frequency distributions on the choice of search parameters, in the next set of experiments we chose the following values of the search parameters: maximum number of experiments is 20, maximum number of moves per experiment is 20, diversifica-

tion frequency is 10 (i.e., after 10 moves take several random moves), and diversification duration is 2 (i.e., take 2 random moves after every 10 moves). Note that this choice of parameters is at the upper limit of reasonable parameter choices, since if a graphlet is very infrequent or does not exist in a graph, we still do 20 experiments with 20 moves each attempting to find the graphlet and all this is done the number of seed node times; i.e., for $|V|$ seed nodes, we make $20 \times 20 \times |V| = 400|V|$ moves looking for a graphlet that is infrequent or does not appear in the network. However, we wanted to determine the effect of these extreme parameter choices on the quality of the obtained results. As before, we experimented with different numbers of seed nodes: for each graph $G(V, E)$ processed by this heuristic, we performed experiments using $|V|/8$, $|V|/4$, $|V|/2$, $|V|$, $2|V|$, $4|V|$, and $8|V|$ seed nodes, respectively. We performed 10 distinct runs of the algorithm for most of the graphs and number of seed nodes; that is, for a graph G, if we ran the algorithm 10 times for each of the 7 seed node numbers, $10 \times 7 = 70$ different runs were performed. The averages and standard deviations of the estimated graphlet frequencies were obtained for each of the 10 runs of the algorithm for the same graph and the same number of seed nodes; the standard deviations were several orders of magnitude smaller than the corresponding averages.

As before, we analyzed the yeast high-confidence PPI network and its corresponding model networks using NLS with this choice of search parameters. We plotted the average graphlet frequencies obtained by these experiments against the exact graphlet frequencies for the PPI and the corresponding model networks (Supplementary Figures 19 and 20). Note that this heuristic approach may result in both under-counting and over-counting of graphlets. Over-counting happens when the parameters are chosen to be too large, i.e., when we search "too hard" for a graphlet, and thus we find the same instance of a graphlet multiple times. Distances between the estimated and the exact results are presented in Supplementary Table 28. As before, the resulting graphlet frequency distributions presented in Supplementary Figures 19 – 20 and Supplementary Table 28 suggest that graphlet counts of PPI and the corresponding geometric random networks are reasonably well approximated by this heuristic, while graphlet counts of ER, ER-DD, and SF networks are not. However, note that with this choice of the search parameters, some graphlets are found as many as NUM-EXP times while others are not (Supplementary Figures 19 – 20). Consequently, the quality of the resulting graphlet frequency distribution estimates goes down: the distances between the exact and the estimated graphlet frequency distributions obtained with these search parameters are higher than those between the exact and the estimated graphlet frequency distributions obtained with the above described lower search parameters for all tested networks and seed node numbers except for the PPI network, for which they are marginally lower (Supplementary Tables 27 and 28). Thus, the previous choice

of search parameters that corresponds to a light random neighborhood local search gives better graphlet frequency distribution estimates.

It is interesting that all of the NLS-estimated graphlet frequencies follow exactly the same pattern. That is, the graphlet frequency distribution estimates in Supplementary Figure 21 and Supplementary Figures 19 and 20 are all "parallel" to each other. They are also "getting closer" to the exact values in PPI and geometric random networks with increasing numbers of samples. Thus, we expect that the results of this algorithm would converge close to the real values in these networks with an increased number of samples. However, since in each of the experiments, including those with the smallest number of samples (i.e., numbers of seed nodes), the same pattern of the graphlet frequency distribution is obtained, we recommend the same strategy as before to obtain convergent results: rather than taking more samples, take very few samples and analytically correct for uniform under-counting of graphlets in PPI and geometric random networks. More experiments are needed to do this with confidence.

The average processing times taken by these experiments are presented in Supplementary Tables 29 and 30. They are much smaller than the exhaustive search processing times for PPI and geometric random networks. For example, a good estimate of the graphlet frequency distribution for the yeast high-confidence PPI network can be achieved in under 6 minutes using this approach (Supplementary Table 30, $\frac{|V|}{8}$ seed nodes), while the exhaustive search takes almost 9 hours; that is, the ratio of the exhaustive search time, T_E and this heuristic search time, T_H, is $r = \frac{T_E}{T_H} = \frac{8h57m16.63s}{5m40.20s} = \frac{32236.63s}{340.20s} = 94.76 \approx 95$. Similarly, this ratio for the tested geometric random networks is as high as 377. However, the processing times of these experiments are much higher for ER, ER-DD, and SF networks when compared to the results of the exhaustive searches. This is due to the algorithm searching for graphlets that are very infrequent, or do not exist at all, in these networks. Since only the sparse graphlets are frequent in these networks, this results in a lot of wasted time as most of the graphlets, i.e. all of the denser ones, are infrequent, or non-existent, in these networks. Thus, this approach should not be used for ER, ER-DD, and SF networks.

As expected, the processing times increase with increased numbers of samples. However, it is interesting that by taking fewer samples we do not loose accuracy of estimated graphlet frequency distribution pattern. (Remember that the results of the TLNP and TNP heuristic approaches behaved this way as well.) Also, with increased dimensionality and density of GEO networks, the processing time grows as a result of larger local neighborhoods having to be explored (the same is true for the TNP heuristic). Note that the slight discrepancy in the processing times for different values of search parameters is due to the large variability in processor

speeds in the CDF machine cluster. We used one CPU from the cluster to do one whole experiment and the CPUs were selected based on their availability and not on their processing speeds. However, the standard deviations of running times of the experiments of the same type were still small compared to the average running times.

We expect this approach to be scalable to large geometric random and PPI networks, since it is based on randomly sampling the network and doing a very limited search in the neighborhood of a randomly chosen graphlet. It is also encouraging that using very few samples yields very good estimates, so it is possible that even as few as $1\%|V|$ or even fewer samples, as was the case in the TLNP and TNP heuristic approaches, will give good estimates of graphlet frequency distributions in PPI and geometric random networks. We hope to experimentally verify these hypotheses in the future. For now, we performed only one set of ten NLS experiments on the 100,000-node, 750,000-edge GEO-3D network described in the previous section. We chose the number of seed nodes to be very small, $\frac{|V|}{1000} = 100$. The resulting estimated graphlet frequency distribution for this network is presented in Supplementary Figure 22; the average of the graphlet frequencies obtained from the 10 experiments is presented. This estimated graphlet frequency distribution is at distance 45.28 from the exact graphlet distribution of a GEO-3D network corresponding to the yeast high-confidence PPI network with 3 times as many edges as the PPI network. The average CPU time over the 10 experiments was 10h 10m 0.27s and the standard deviation was 1h 45m 19.39s. This larger running time is not only due to the increased graph density, but also to the nature of the LEDA library in which we first load the graph and then need to do linear time enumeration of its nodes and edges. An improved implementation that would do the node and edge enumeration (and storage in data structures) required for random sampling as the graph is being loaded would result in a decrease in this running time.

To conclude, the NLS-based heuristics graphlet search approach gives good graphlet frequency distribution estimates for PPI and geometric random networks. However, it does not work well for Erdös-Rényi and Scale-Free networks. Using very few samples yields very good graphlet frequency distribution estimates for PPI and geometric random networks. The experiments with low values of search parameters give better approximations than those with high values of these parameters.

4 Conclusions and Future Work

We have described two heuristic graphlet frequency estimation approaches that work well for PPI and geometric random networks. They do not work well for

ER, ER-DD, and SF random networks both in terms of the resulting estimates and running times. Note that both of these approaches work well for PPI networks, which have scale-free degree distributions and contain hubs, and also for geometric random networks, which have Poisson degree distributions and lack hubs. Thus, it is not the presence or absence of hubs that dictates the behavior of these heuristics, as was the case in the Kashtan et al. (2004) algorithm, but the local structure of the networks. Surprisingly small samples were needed to produce very good estimates of graphlet frequency distribution patterns in PPI and geometric random networks.

Even though we have obtained excellent relative graphlet frequency estimates, more experimentation is needed in order to determine approaches that would "translate" the estimated graphlet frequency distributions closer to the exact one in absolute values. Also, a more detailed theoretical explanation of the relationship between the structure of the networks and the success of the heuristic approaches would be beneficial.

5 Supplementary Figures

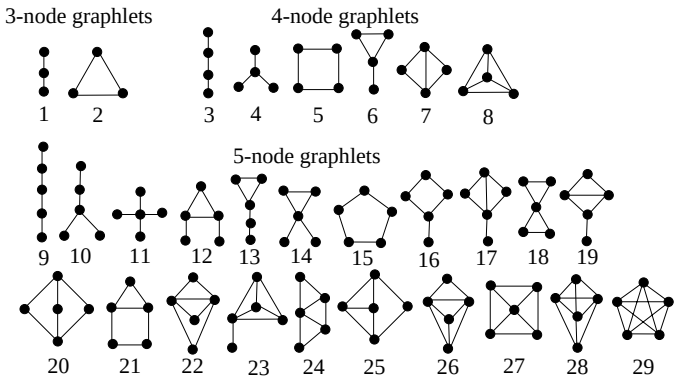


Figure 1:
All 3-node, 4-node, and 5-node connected networks (graphlets), ordered within groups from the least to the most dense with respect to the number of edges when compared to the maximum possible number of edges in the graphlet; they are numbered from 1 to 29 (Pržulj et al. 2004).

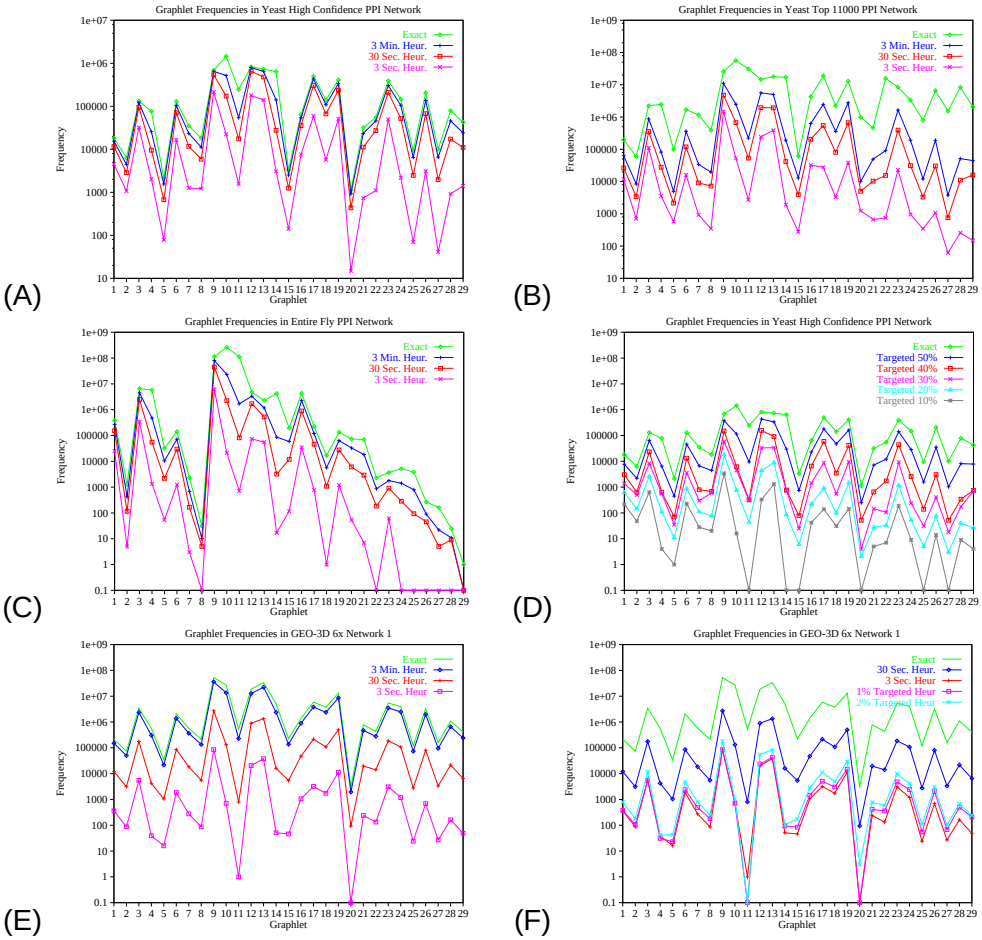


Figure 2: Comparison of exact graphlet frequencies in the high-confidence and top 11000 *S. cerevisiae* PPI network (von Mering et al. 2002), the entire currently available noisy *D. melanogaster* PPI network (Giot et al. 2003), and a geometric random network (green line) with the corresponding TLNP and TNP heuristic graphlet frequency estimates. Zero frequencies were replaced by 0.1 for plotting on log-scale. **A.** TLNP estimates for high-confidence *S. cerevisiae* PPI network. **B.** TLNP estimates for the top 11000 *S. cerevisiae* PPI network. **C.** TLNP estimates for the noisy *D. melanogaster* PPI network. **D.** TNP estimates for high-confidence *S. cerevisiae* PPI network. **E.** TLNP estimates for a GEO-3D network. **F.** TLNP and TNP estimates for a GEO-3D network.

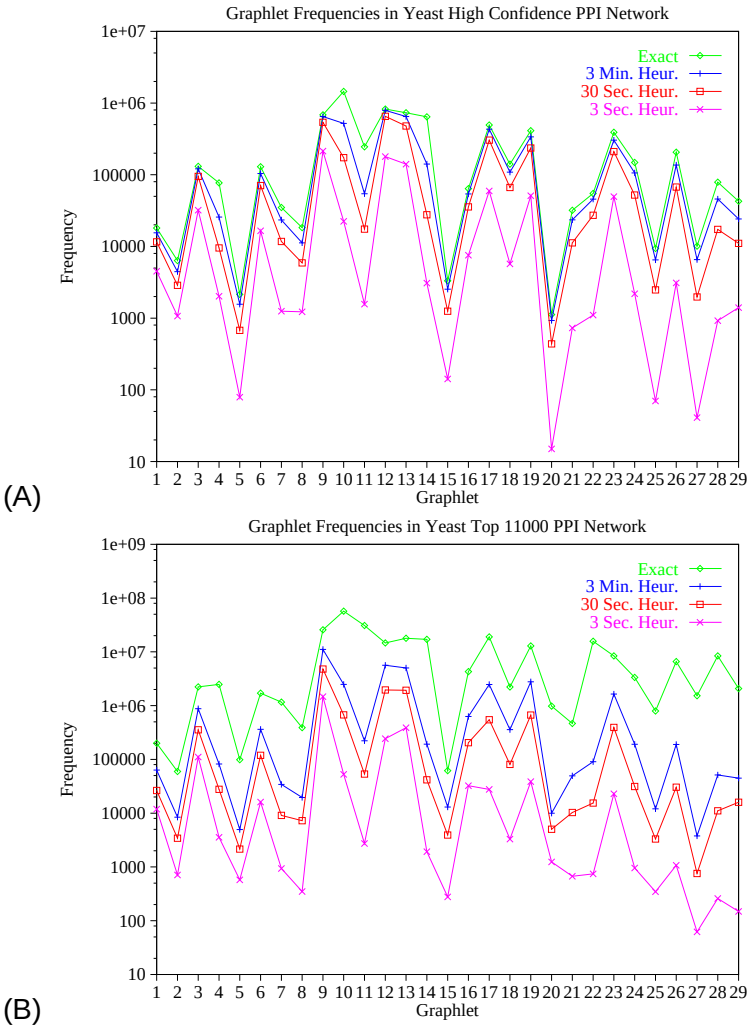


Figure 3:
Comparison of the exact graphlet counts and those obtained by the TLNP heuristic graphlet searches for yeast PPI networks: **A.** Yeast high confidence PPI network; **B.** Yeast top 11000 PPI network.

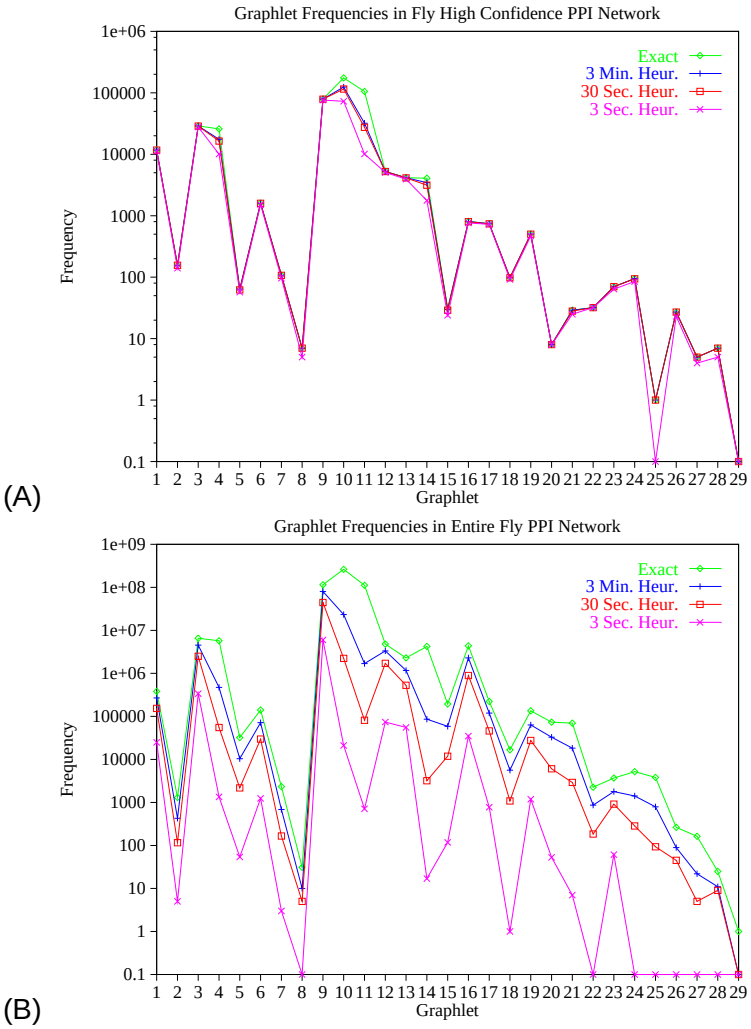


Figure 4:
Comparison of the exact graphlet numbers and those obtained by the TLNP heuristic graphlet searches for fruitfly PPI networks: **A.** High confidence fruitfly PPI network; **B.** Entire fruitfly PPI network.

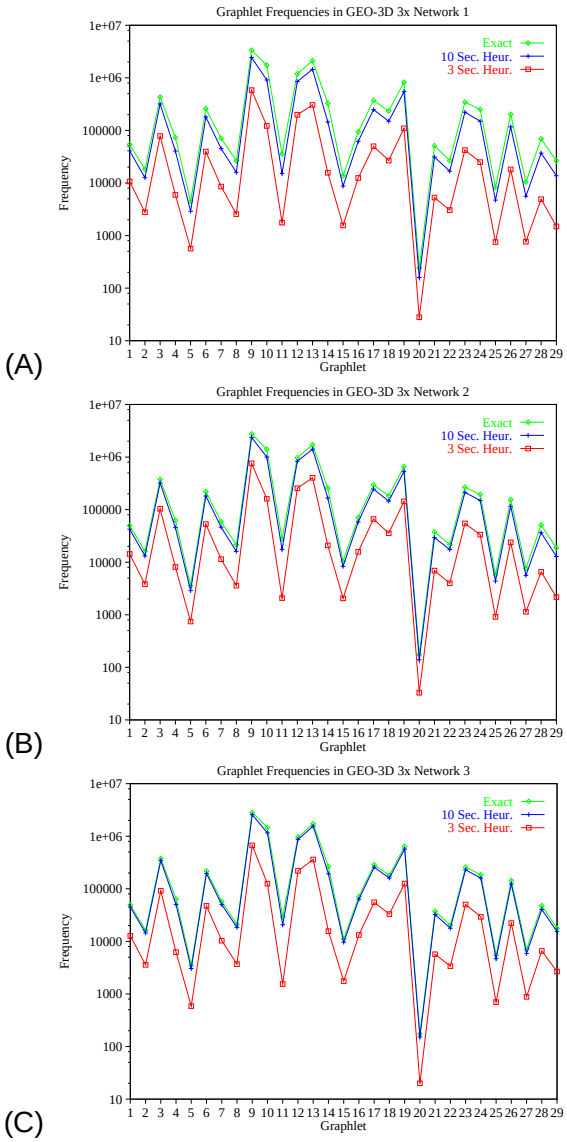


Figure 5:
Comparison of the exact and the TLNP heuristic graphlet frequency distributions with various processing time cut-offs for the GEO-3D networks corresponding to the yeast high confidence PPI network with 3 times as many edges as the PPI network. **A.** Network 1. **B.** Network 2. **C.** Network 3.

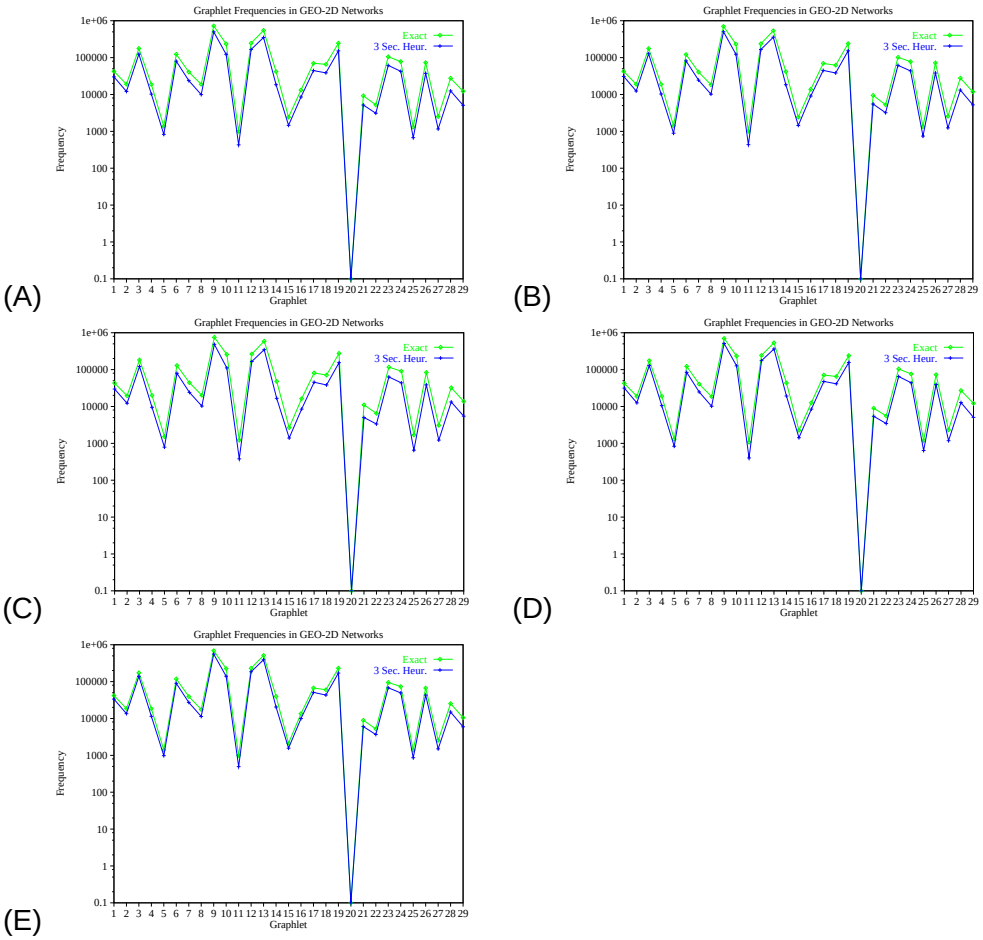


Figure 6:
Comparison of the exact and the TLNP heuristic graphlet frequency distributions for the GEO-2D networks corresponding to the yeast top 11000 PPI network: **A.** Network 1. **B.** Network 2. **C.** Network 3. **D.** Network 4. **E.** Network 5.

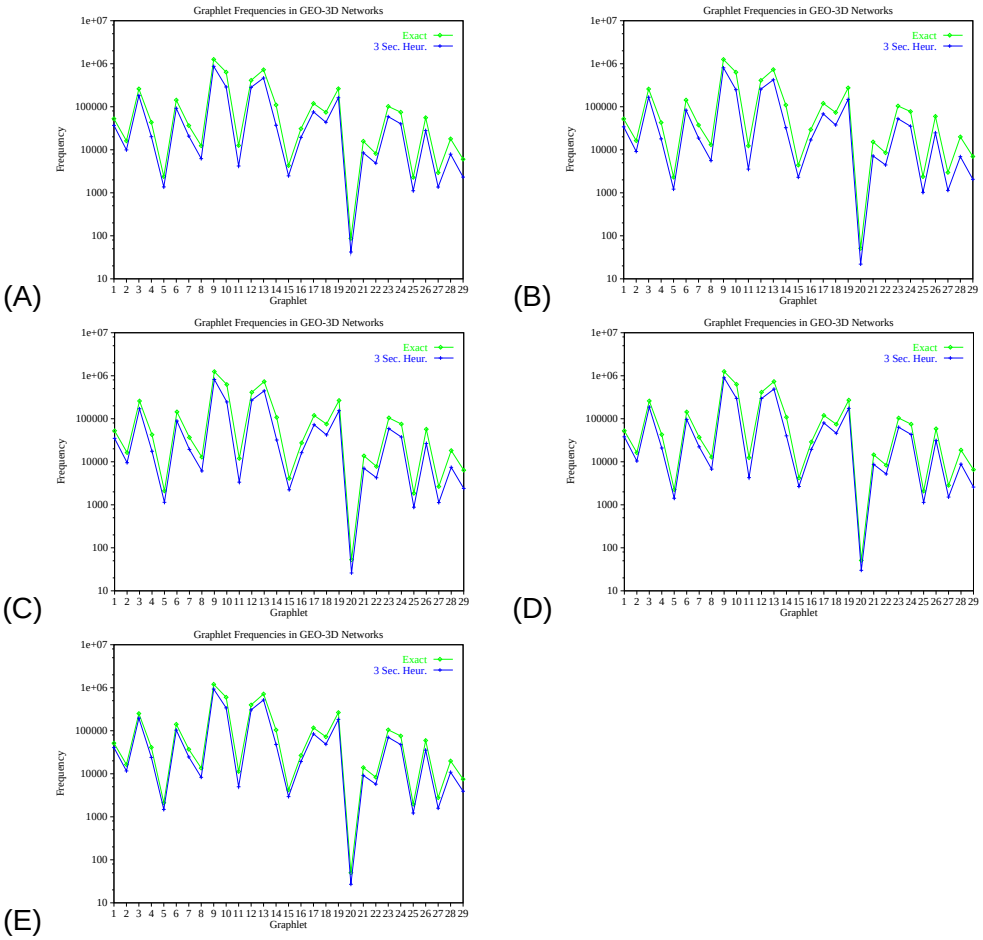


Figure 7:
Comparison of the exact and the TLNP heuristic graphlet frequency distributions for the GEO-3D networks corresponding to the yeast top 11000 PPI network: **A.** Network 1. **B.** Network 2. **C.** Network 3. **D.** Network 4. **E.** Network 5.

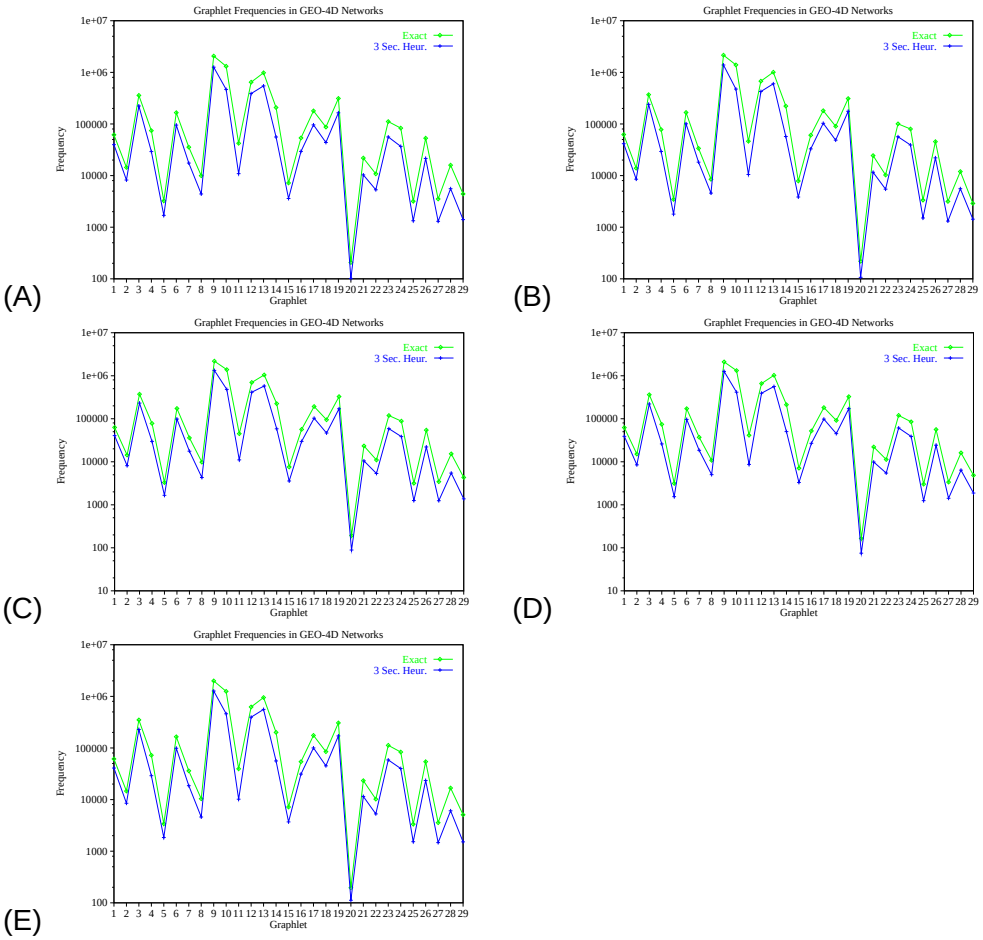


Figure 8:
Comparison of the exact and the TLNP heuristic graphlet frequency distributions for the GEO-4D networks corresponding to the yeast top 11000 PPI network: **A.** Network 1. **B.** Network 2. **C.** Network 3. **D.** Network 4. **E.** Network 5.

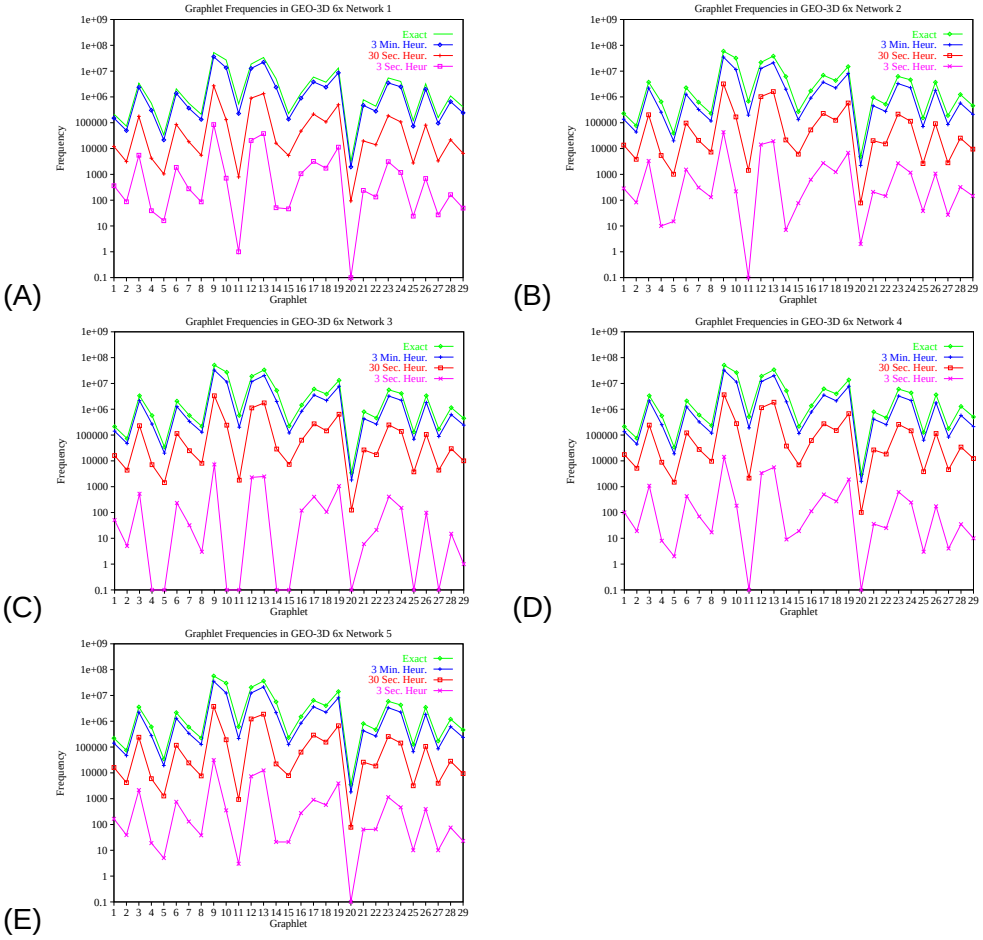
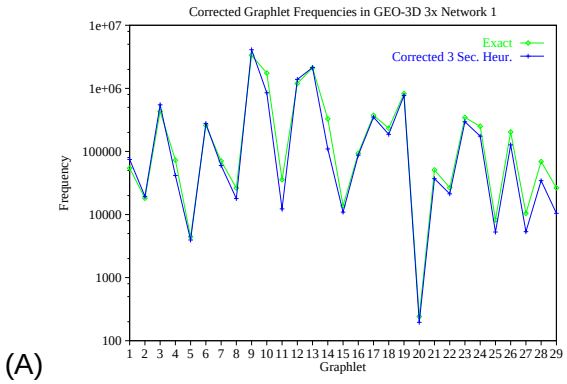
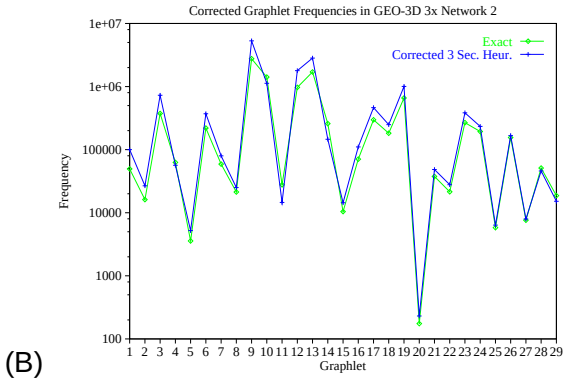


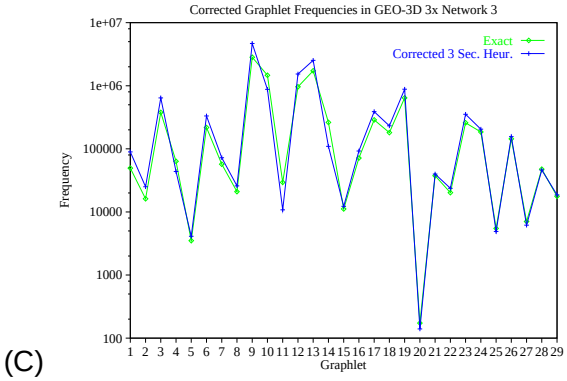
Figure 9:
Comparison of the exact and the TLNP heuristic graphlet searches with various processing time cut-offs for the GEO-3D networks corresponding to the yeast high confidence PPI network with 6 times as many edges as the PPI network. **A.** Network 1. **B.** Network 2. **C.** Network 3. **D.** Network 4. **E.** Network 5.



(A)



(B)



(C)

Figure 10:
Comparison of the exact and the corrected, 3 second time bounded TLNP estimated graphlet frequency distributions for the GEO-3D networks corresponding to the yeast high confidence PPI network with 3 times the number of edges as the PPI network. **A.** Network 1. **B.** Network 2. **C.** Network 3.

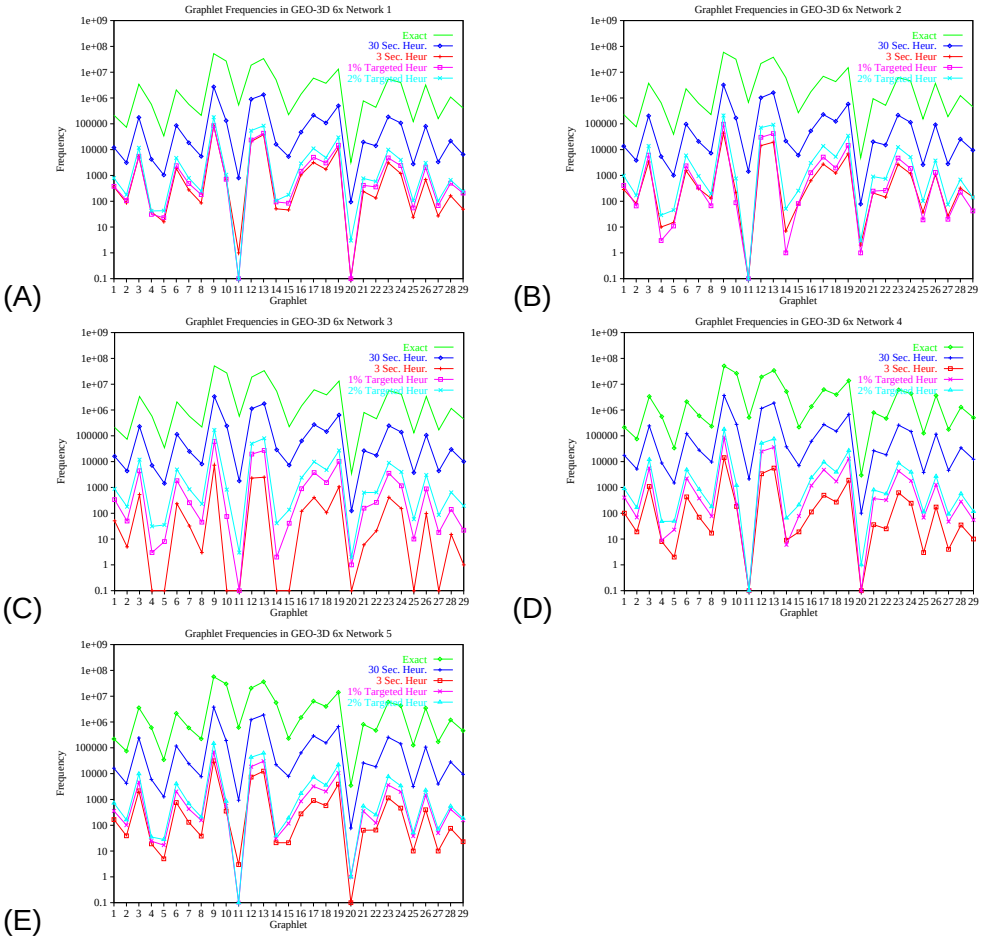


Figure 11:

Comparison of the exact, the TLNP with various processing time cut-offs, and the TNP with different percentages of selected targeted nodes, estimated graphlet distributions, for GEO-3D networks corresponding to the yeast high confidence PPI network with 6 times as many edges as the PPI network. **A.** Network 1. **B.** Network 2. **C.** Network 3. **D.** Network 4. **E.** Network 5.

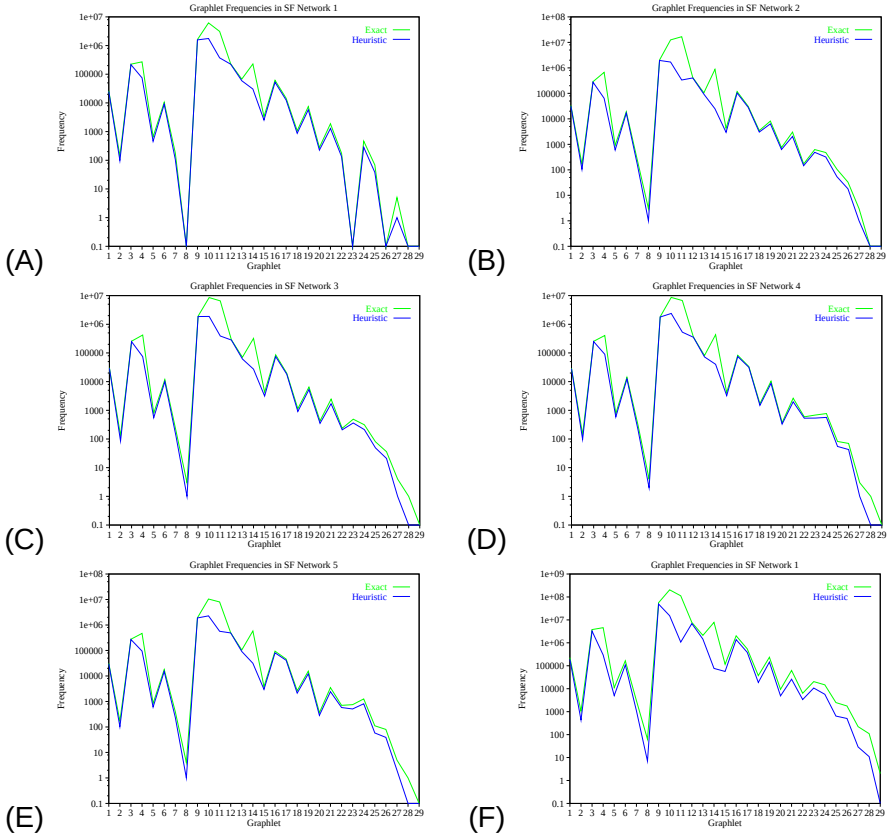


Figure 12:
Comparison of the exact and the TLNP 3 minute time limited heuristic graphlet search results for SF networks corresponding to the yeast high-confidence PPI networks (A-E), and the yeast top 11000 PPI network (F).

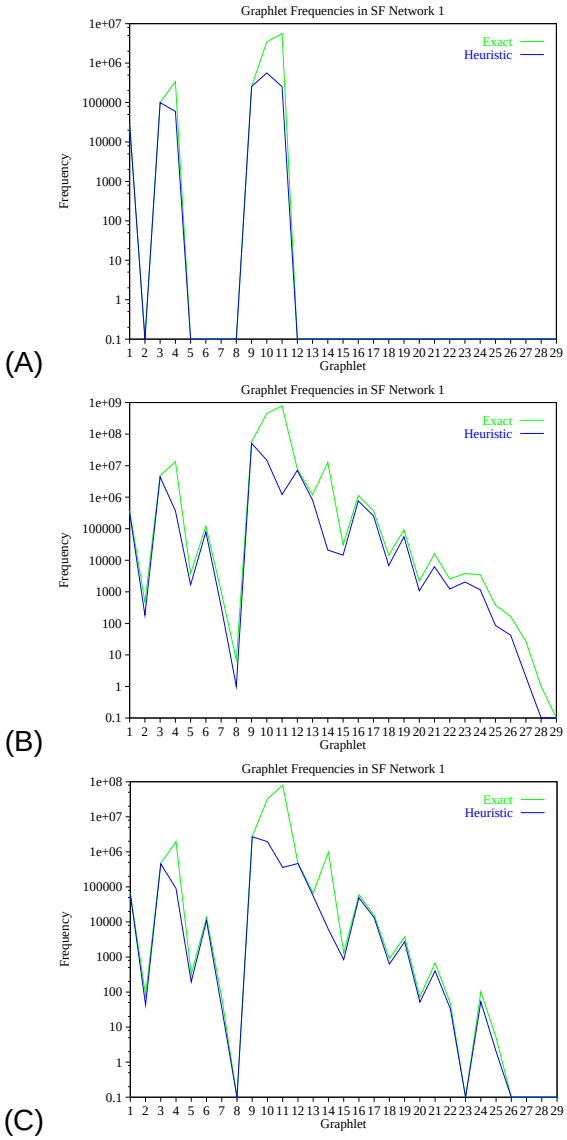


Figure 13:
Comparison of the exact and the TLNP 3 minute time limited heuristic graphlet search results for scale-free networks corresponding to the fruitfly and worm PPI networks: **A.** The high confidence fruitfly PPI network; **B.** The entire fruitfly PPI network; **C.** The entire worm PPI network.

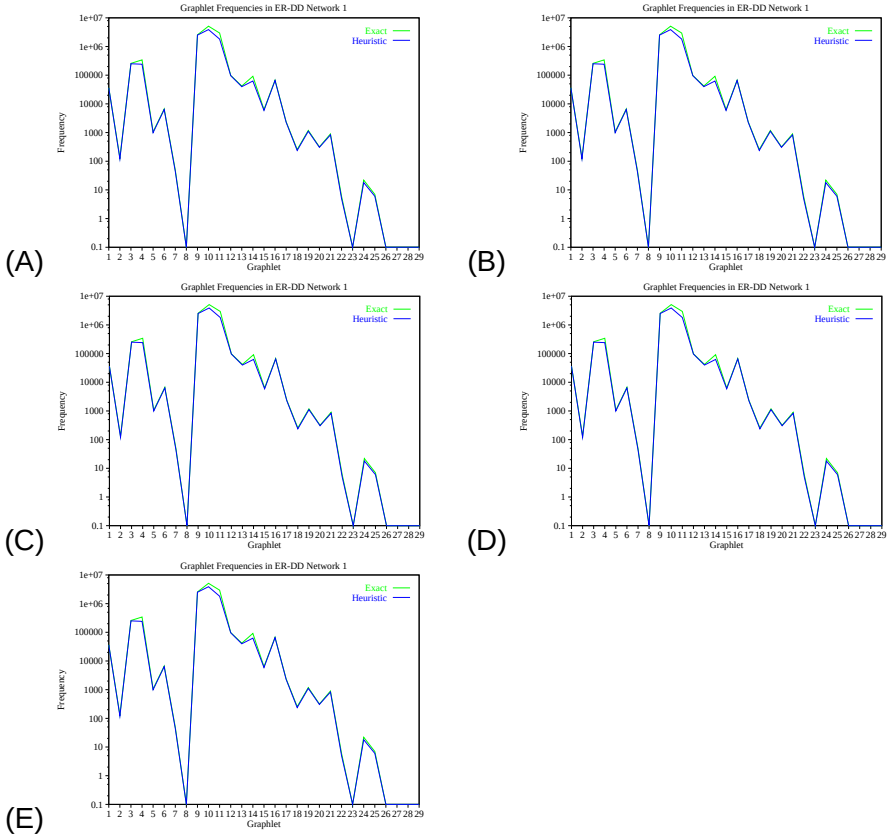


Figure 14:
Comparison of the exact and the TLNP 3 minute time limited heuristic graphlet search results for ER-DD networks corresponding to the yeast high confidence PPI network. **A.** ER-DD network 1. **B.** ER-DD network 2. **C.** ER-DD network 3. **D.** ER-DD network 4. **E.** ER-DD network 5.

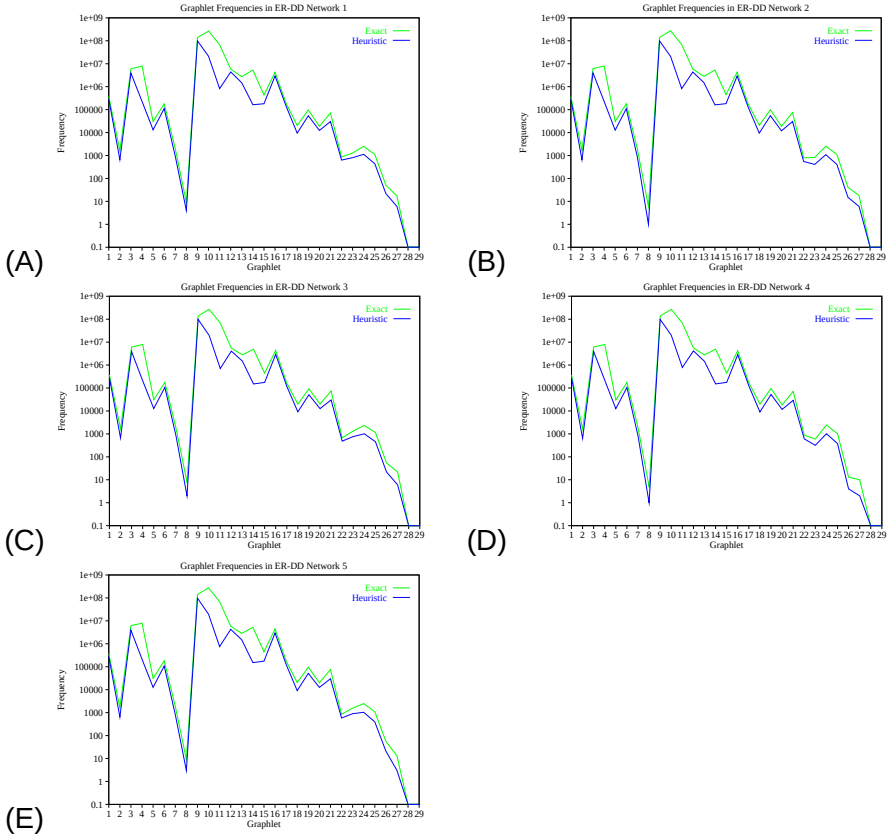
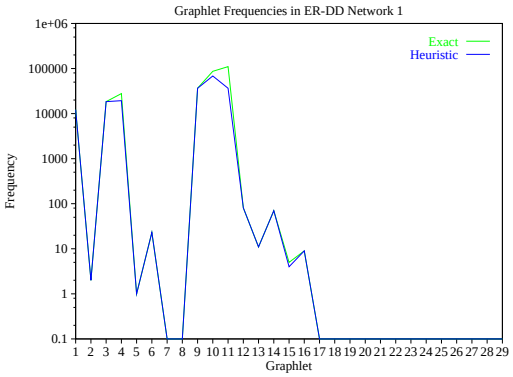
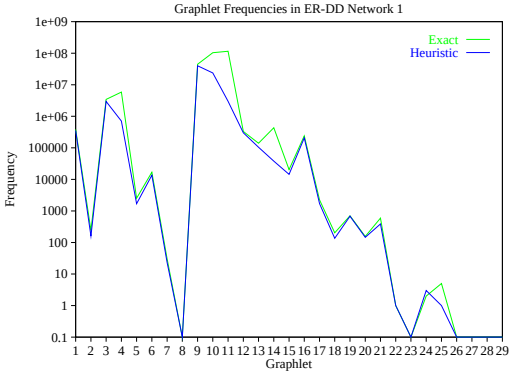


Figure 15:
Comparison of the exact and the TLNP 3 minute time limited heuristic graphlet search results for five ER-DD networks corresponding to the yeast top 11000 PPI network. **A.** ER-DD network 1. **B.** ER-DD network 2. **C.** ER-DD network 3. **D.** ER-DD network 4. **E.** ER-DD network 5.

(A)



(B)



(C)

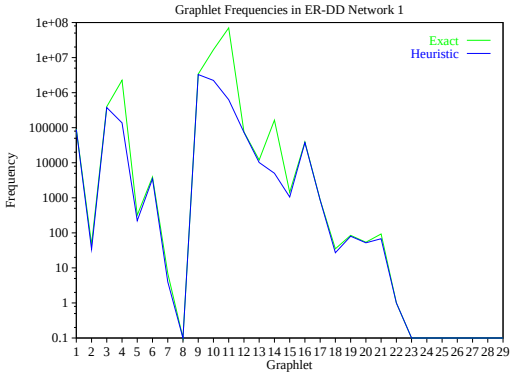


Figure 16:

Comparison of the exact and the TLNP 3 minute time limited heuristic graphlet search results for ER-DD networks corresponding to the fruitfly and worm PPI networks. **A.** ER-DD network corresponding to the high confidence fruitfly PPI network. **B.** ER-DD network corresponding to the entire fruitfly PPI network. **C.** ER-DD network corresponding to the entire worm PPI network.

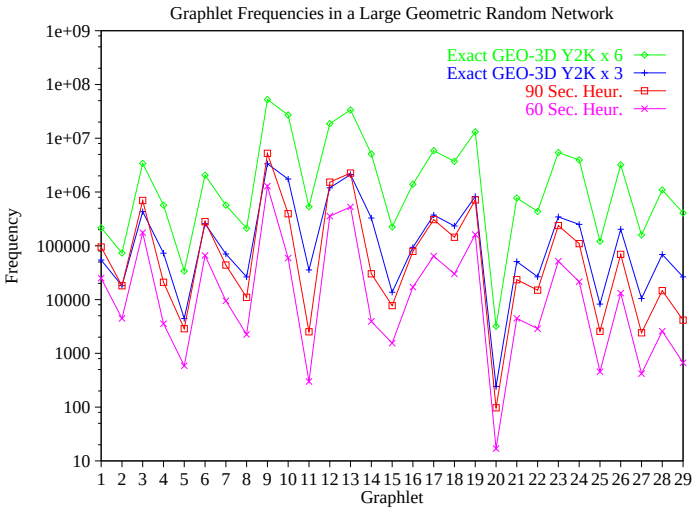


Figure 17:
Graphlet distributions obtained by TLNP with 60 and 90 second node processing cut-off time for a GEO-3D network with 100,000 nodes and 750,000 edges; comparison with the exact graphlet distributions for GEO-3D networks corresponding to yeast high-confidence PPI network with 3 and 6 times as many edges as the PPI network.

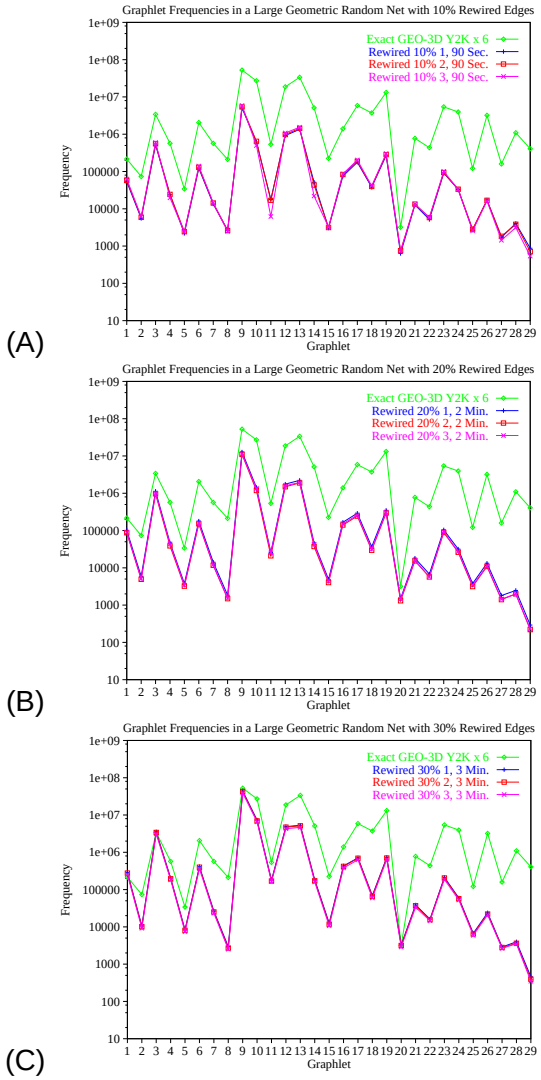


Figure 18:
TLNP estimated graphlet distributions with different node processing cut-off times for GEO-3D networks with 100,000 nodes and 750,000 edges and 10% (A), 20% (B), and 30% (C) of its edges rewired at random. Comparison with the exact graphlet distributions for GEO-3D networks corresponding to the yeast high confidence PPI network with 6 times as many edges as the PPI network.

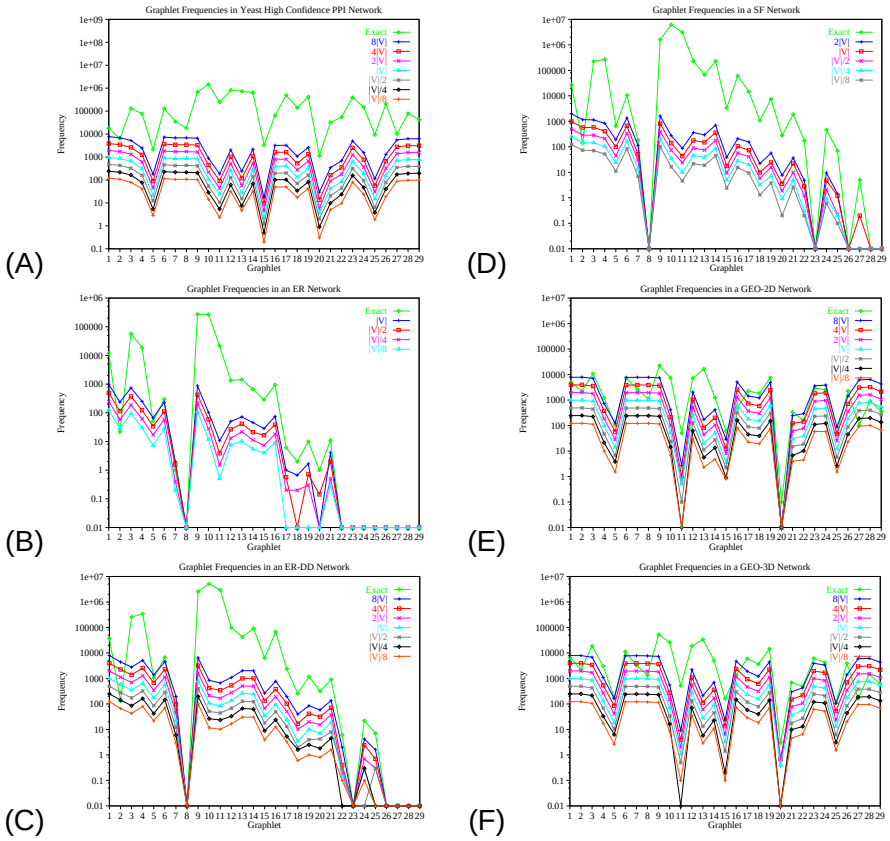


Figure 19:

Exact graphlet frequencies and average graphlet frequencies obtained from a number of runs of the NLS heuristic algorithm with search parameters 20, 20, 10, 2 for the following networks (since some of these averages were close to 0.1, we replaced 0 graphlet frequencies by 0.01 frequencies to obtain nicer looking log-scale plots): **A.** The yeast high-confidence PPI network. Averages are taken for 10 runs for each seed node size. **B.** ER model network corresponding to the yeast high-confidence PPI network. Averages are taken for 3 runs for $|V|$ seed nodes, 7 runs for $|V|/2$, and 10 runs for the other numbers of seed nodes. **C.** ER-DD network corresponding to the yeast high-confidence PPI network. Averages are taken for 5 runs for $8|V|$, and 10 runs for the other numbers of seed nodes. **D.** SF model network corresponding to the yeast high-confidence PPI network. Averages are taken of 7 runs with $2|V|$ seed nodes, 5 runs with $|V|$ seed nodes, 9 runs with $|V|/2$ seed nodes, and 10 runs with $|V|/4$ and $|V|/8$ seed nodes. **E.** GEO-2D model network corresponding to the yeast high-confidence PPI network. Averages are taken for 10 runs for each seed node size. **F.** GEO-3D model network corresponding to the yeast high-confidence PPI network. Averages are taken for 10 runs for each seed node size.

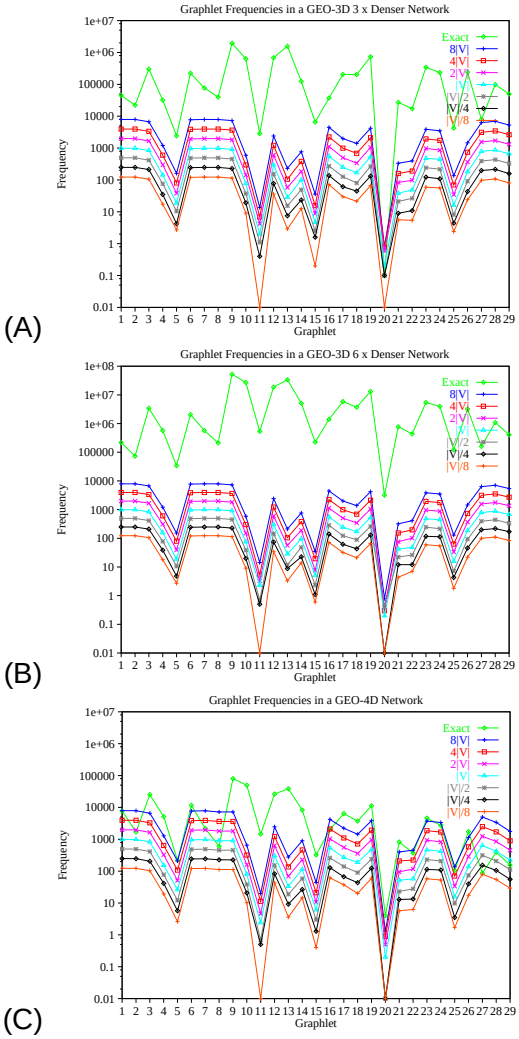


Figure 20:

Exact graphlet frequencies and the average graphlet frequencies obtained from 10 runs of the NLS-based heuristic search algorithms with search parameters 20, 20, 10, 2 for each of the following model networks corresponding to the yeast high-confidence PPI network: **A.** GEO-3D which is 3 times denser than the PPI network. **B.** GEO-3D which is 6 times denser than the PPI network. **C.** GEO-4D. Since some of these averages were close to 0.1, we replaced 0 graphlet frequencies by 0.01 frequencies to obtain nicer looking log-scale plots.

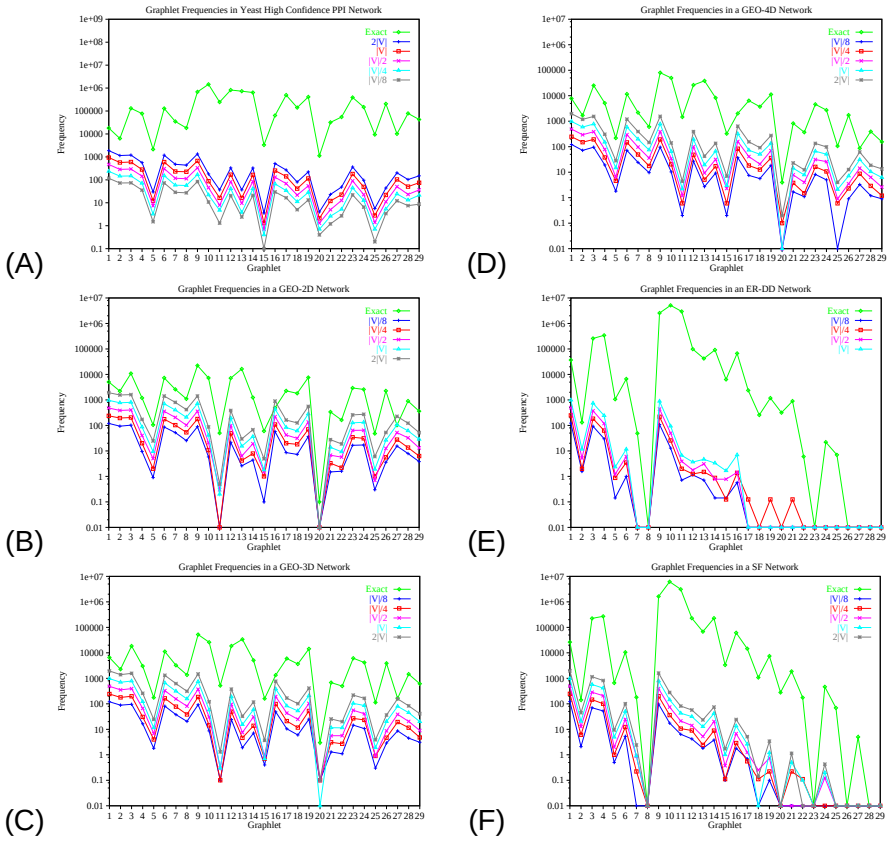


Figure 21: Exact graphlet frequencies and the average graphlet frequencies obtained from 10 runs of the NLS heuristic graphlet search algorithm with search parameters 2, 5, 3, 1 and $2|V|$, $|V|$, $|V|/2$, $|V|/4$, and $|V|/8$ seed nodes per n-node, m-edge graphlet, for each of the following networks (since some of these averages were close to 0.1, we replaced 0 graphlet frequencies by 0.01 frequencies to obtain nicer looking log-scale plots): **A.** The yeast high-confidence PPI network. **B.** A GEO-2D model network corresponding to the yeast high-confidence PPI network. **C.** A GEO-3D model network corresponding to the yeast high-confidence PPI network. **D.** A GEO-4D model network corresponding to the yeast high-confidence PPI network. **E.** An ER-DD model network corresponding to the yeast high-confidence PPI network. **F.** An SF model network corresponding to the yeast high-confidence PPI network.

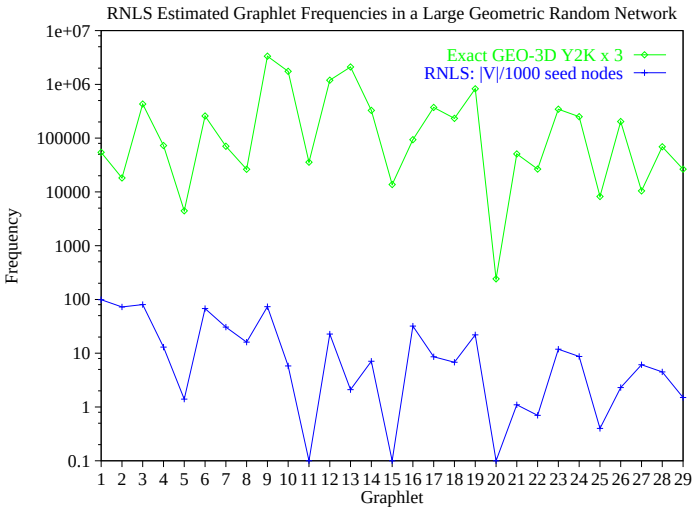


Figure 22:
NLS estimated heuristic graphlet frequency distribution with 100 seed nodes for a GEO-3D network with 100,000 nodes and 750,000 edges. The resulting estimates are averages over 10 different NLS experiments. Comparison is done with the exact graphlet distribution of the GEO-3D network corresponding to the yeast high-confidence PPI network with 3 times as many edges as the PPI network. These two networks are of similar edge densities.

6 Supplementary Tables

		Nodes ordered by eccentricity and degree																		
node number in order:		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Graph 1	eccentricity	10	10	10	10	10	10	9	9	9	9	9	9	9	9	9	9	9	9	9
	degree	5	10	11	11	14	16	5	5	6	6	7	7	8	8	9	9	9	9	9
Graph 2	eccentricity	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
	degree	7	7	8	8	8	8	8	9	9	9	9	9	9	10	10	10	10	11	11
Graph 3	eccentricity	10	10	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
	degree	5	8	4	6	6	7	8	8	9	10	10	10	10	11	11	11	11	11	11
Graph 4	eccentricity	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
	degree	7	7	8	9	9	9	9	9	9	10	10	10	10	10	10	10	10	10	11
Graph 5	eccentricity	10	10	10	10	10	10	10	9	9	9	9	9	9	9	9	9	9	9	9
	degree	4	5	8	12	13	14	16	4	5	6	7	7	7	7	8	8	8	8	9

Table 1: Degrees and eccentricities of the selected 2% nodes to be processed in the five 3-dimensional geometric random graphs with the same number of nodes and six times as many edges as the yeast high-confidence PPI network.

	3 minute			30 second			3 second		
Net.	num.	%	D	num.	%	D	num.	%	D
YHC	14	1.4%	7.58	36	3.6%	15.14	114	11.5%	37.94
Y11K	250	10.4%	37.86	430	17.9%	42.89	920	38.3%	68.44
FH	1	0.02%	9.95	2	0.04%	11.14	24	0.5%	17.90
FE	331	4.7%	22.35	1104	15.8%	33.19	3961	56.7%	83.69

Table 2: Number of unprocessed nodes in the five PPI networks by the TLNP heuristic search algorithm and distances between the exact and heuristic graphlet counts. "YHC" denotes the yeast high-confidence, "Y11K" the yeast top 11000, "FH" the fruitfly high-confidence, and "FE" the entire fruitfly. Columns denoted by "num." present the number of unfinished nodes for the 3 minute, 30 second, and 3 second cut-off times while columns denoted by "%" present fractions of nodes, with respect to the total number of nodes in a network, which remained unprocessed by the heuristic. Columns denoted by "D" show distances between the heuristic and exact graphlet counts.N/A denotes that experiments for the given network and time bound are not needed and, thus, were not performed.

	Yeast High-Confidence			Yeast Top 11000		
Graphlet	3m	30s	3s	3m	30s	3s
1	0.8563	0.6409	0.2493	0.3164	0.1321	0.0595
2	0.7016	0.4519	0.1680	0.1414	0.0574	0.0119
3	0.9220	0.7223	0.2447	0.3932	0.1589	0.0497
4	**0.3341**	**0.1241**	0.0263	0.0330	0.0112	0.0014
5	0.7277	0.3156	0.0368	0.0503	0.0218	0.0058
6	0.8046	0.5470	0.1280	0.2129	0.0700	0.0095
7	0.6697	0.3351	0.0357	0.0295	0.0078	0.0008
8	0.6122	0.3218	0.0670	0.0502	0.0186	0.0009
9	0.9478	0.7876	0.3119	0.4302	0.1861	0.0570
10	**0.3585**	**0.1191**	0.0154	0.0435	0.0118	0.0009
11	**0.2216**	**0.0711**	**0.0064**	**0.0072**	**0.0017**	**0.0001**
12	0.9590	0.7914	0.2178	0.3848	0.1335	0.0166
13	0.8878	0.6554	0.1924	0.2805	0.1081	0.0217
14	**0.2200**	**0.0431**	**0.0048**	**0.0112**	**0.0024**	**0.0001**
15	0.7655	0.3769	0.0428	0.2108	0.0636	0.0045
16	0.8436	0.5568	0.1182	0.1466	0.0477	0.0076
17	0.8727	0.6141	0.1203	0.1302	0.0287	0.0015
18	0.7735	0.4739	0.0407	0.1612	0.0363	0.0015
19	0.8179	0.5722	0.1234	0.2160	0.0518	0.0030
20	0.8262	0.3904	**0.0134**	**0.0102**	0.0051	0.0013
21	0.7347	0.3511	0.0228	0.1060	0.0221	0.0014
22	0.8333	0.4963	0.0201	**0.0058**	**0.0010**	**0.0000**
23	0.7766	0.5373	0.1266	0.1950	0.0469	0.0027
24	0.7140	0.3509	0.0147	0.0571	0.0094	0.0003
25	0.6929	0.2644	**0.0075**	0.0151	**0.0042**	0.0004
26	0.6624	0.3263	0.0151	0.0287	0.0046	0.0002
27	0.6488	**0.1955**	**0.0041**	**0.0025**	**0.0005**	**0.0000**
28	**0.5826**	**0.2196**	**0.0117**	**0.0061**	**0.0013**	**0.0000**
29	**0.5656**	0.2580	0.0329	0.0216	0.0077	**0.0001**

Table 3: Fraction between the TLNP heuristic estimate of the number of graphlets of type i and the exact number of graphlets of type i, $1 \leq i \leq 29$ for the two yeast PPI networks. The TLNP heuristic graphlet search algorithm was run with three different node processing time cut-offs. "3s", "30s", and "3m" denote 3 second, 30 second, and 3 minute time cut-offs, respectively. In each column, the smallest 6 numbers are presented in bold.

	Fly High-Confidence			Fly Entire		
Graphlet	3m	30s	3s	3m	30s	3s
1	0.9967	0.9943	0.9641	0.7050	0.4020	0.0652
2	**0.9936**	0.9936	0.8917	0.3300	0.0892	0.0038
3	0.9962	0.9932	0.9599	0.6898	0.3805	0.0516
4	**0.6760**	**0.6334**	**0.3876**	**0.0826**	**0.0096**	0.0002
5	1.0000	0.9688	0.8906	0.3227	0.0677	0.0017
6	0.9975	0.9938	0.9419	0.5080	0.2106	0.0088
7	1.0000	0.9907	0.8889	0.2937	0.0713	0.0013
8	1.0000	1.0000	0.7143	0.3226	0.1613	0.0032
9	0.9965	0.9936	0.9602	0.6994	0.3872	0.0520
10	**0.7088**	**0.6575**	**0.4146**	**0.0889**	**0.0085**	0.0001
11	**0.3042**	**0.2610**	**0.0961**	**0.0151**	**0.0007**	**0.0000**
12	0.9979	0.9962	0.9640	0.6929	0.3559	0.0154
13	0.9976	0.9906	0.9444	0.5115	0.2315	0.0243
14	**0.8454**	**0.7709**	**0.4342**	**0.0203**	**0.0008**	**0.0000**
15	0.9677	**0.9355**	0.7742	0.3031	0.0611	0.0006
16	0.9975	0.9889	0.9443	0.5258	0.2041	0.0080
17	1.0000	0.9960	0.9784	0.5331	0.2062	0.0035
18	1.0000	1.0000	0.9388	0.3370	0.0647	0.0001
19	0.9960	0.9921	0.9267	0.4700	0.2041	0.0088
20	1.0000	1.0000	1.0000	0.4470	0.0825	0.0007
21	1.0000	0.9655	0.8621	0.2625	0.0418	0.0001
22	1.0000	1.0000	1.0000	0.3834	0.0815	**0.0000**
23	1.0000	1.0000	0.9143	0.4856	0.2473	0.0166
24	1.0000	0.9895	0.8947	0.2719	0.0547	**0.0000**
25	1.0000	1.0000	**0.1000**	0.2077	**0.0245**	**0.0000**
26	1.0000	1.0000	0.8519	0.3384	0.1711	0.0004
27	1.0000	1.0000	0.8000	0.1350	0.0307	0.0006
28	1.0000	1.0000	0.7143	0.4400	0.3600	0.0040
29	1.0000	1.0000	1.0000	**0.1000**	0.1000	0.1000

Table 4: Fraction between the TLNP heuristic estimate of the number of graphlets of type i and the exact number of graphlets of type i, $1 \leq i \leq 29$ for the two fruitfly PPI networks. The TLNP heuristic graphlet search algorithm was run with three different node processing time cut-offs. "3s", "30s", and "3m" denote 3 second, 30 second, and 3 minute time cut-offs, respectively. In each column, the smallest 5 numbers are presented in bold.

	Largest Connected Component		Degree			
Network G	Radius	Diameter	min.	max.	avg.	stDev.
YHC	7	14	1	51	4.97	7.45
Y11K	8	15	1	114	9.16	15.52
FH	14	27	1	38	2.02	1.81
FE	6	11	1	173	5.73	9.12

Table 5: PPI network statistics. The second and third columns present radii and diameters of the largest connected components of the PPI networks, while other columns give degree statistics of the entire network: minimum degree, maximum degree, average degree, and the standard deviation of the degrees, respectively. "YHC" denotes the yeast high confidence PPI network, "Y11K" denotes the yeast top 11000 PPI network, "FH" denotes the fruitfly high confidence PPI network, "FE" denotes the fruitfly entire PPI network.

		Eccentricity				Degree			
Network	num.	min.	max.	avg.	stDev.	min.	max.	avg.	stDev.
YHC 3 sec.	114	7	9	8.42	0.56	2	51	20.93	11.86
YHC 30 sec.	36	7	9	8.06	0.33	21	51	34.89	9.82
YHC 3 min.	14	7	8	7.93	0.27	38	51	44.71	4.50
Y11K 3 sec.	920	8	11	9.92	0.68	1	114	19.83	20.92
Y11K 30 sec.	430	8	11	9.88	0.60	1	114	32.54	24.61
Y11K 3 min.	250	8	11	9.93	0.51	5	114	44.36	25.82
FH 3 sec.	24	14	19	16.50	1.38	10	38	14.25	5.68
FH 30 sec.	2	14	17	15.50	2.12	20	38	29.00	12.73
FH 3 min.	1	14	14	14.00	N/A	38	38	38.00	N/A
FE 3 sec.	3961	6	9	7.60	0.49	1	173	8.93	11.06
FE 30 sec.	1104	6	8	7.10	0.31	3	173	21.13	14.66
FE 3 min.	331	6	8	6.99	0.16	3	173	36.58	18.07

Table 6: Eccentricity and degree statistics for nodes in PPI networks that remained unfinished by the TLNP with given time bounds. The second column, denoted by "num.", gives the number of unfinished nodes, while "min", "max", "avg." and "stDev" denote minimum, maximum, average, and standard deviation of eccentricities and degrees, respectively.

% Finished	CPU Time	Distance
10%	6s	68.34
20%	47s	45.91
30%	3m 40s	39.37
40%	11m 13s	39.80
50%	66m 19s	16.59
exhaustive (100%)	8h 57m 17s	0

Table 7: Percentages of processed nodes, processing times, and graphlet distribution distances of the TNP heuristic graphlet counts from the exact graphlet counts for the yeast high-confidence PPI network.

	3 minute			30 second			10 second			3 second		
Graph	num.	%	D	num.	%	D	num.	%	D	num.	%	D
GEO-2D 1	0	0%	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GEO-2D 2	0	0%	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GEO-2D 3	0	0%	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GEO-2D 4	0	0%	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GEO-2D 5	0	0%	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GEO-3D 1	0	0%	0	N/A	N/A	N/A	N/A	N/A	N/A	3	0.3%	0.78
GEO-3D 2	0	0%	0	0	0%	0	0%	0	0	0	0%	0
GEO-3D 3	0	0%	0	0	0%	0	0%	0	0	0	0%	0
GEO-3D 4	0	0%	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GEO-3D 5	0	0%	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GEO-4D 1	0	0%	0	0	0%	0	0	0%	0	0	0%	0
GEO-4D 2	0	0%	0	0	0%	0	0	0%	0	0	0%	0
GEO-4D 3	0	0%	0	0	0%	0	0	0%	0	0	0%	0
GEO-4D 4	0	0%	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GEO-4D 5	0	0%	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GEO-3D-3x 1	1	0.1%	0.01	18	1.82%	0.61	139	14.07%	3.19	661	66.90%	9.68
GEO-3D-3x 2	0	0%	0	4	0.40%	0.10	77	7.79%	1.84	547	55.36%	8.29
GEO-3D-3x 3	0	0%	0	2	0.20%	0.07	50	5.06%	1.05	575	58.20%	8.49
GEO-3D-3x 4	0	0%	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GEO-3D-3x 5	0	0%	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
GEO-3D-6x 1	198	20.04%	2.11	843	88.09%	16.29	N/A	N/A	N/A	977	98.89%	40.60
GEO-3D-6x 2	258	26.11%	3.43	830	84.00%	17.84	N/A	N/A	N/A	981	99.29%	32.76
GEO-3D-6x 3	198	20.04%	2.90	814	82.39%	14.94	N/A	N/A	N/A	986	99.80%	67.22
GEO-3D-6x 4	223	23.58%	3.52	788	79.76%	14.80	N/A	N/A	N/A	985	99.70%	39.02
GEO-3D-6x 5	219	22.16%	2.72	810	81.98%	17.48	N/A	N/A	N/A	981	99.29%	35.99

Table 8: The number of unprocessed nodes by the TLNP heuristic search algorithm with different node processing cut-off times in the geometric random networks corresponding to the yeast high confidence PPI network and the distances from the exact graphlet counts. Columns denoted by "num." present number of nodes unfinished by the TLNP with 3 minute, 30 second, 10 second, and 3 second cut-off times, respectively, while columns denoted by "%" present fractions of unprocessed nodes by these heuristic experiments with respect to the total number of nodes in the corresponding network. Columns denoted by "D" present the distance of the heuristic graphlet distribution from the exact one. "N/A" indicates that the experiments for the given network and the time bound were not needed, or were not performed. As before, "GEO-2D", "GEO-3D", and "GEO-4D" denote the 2-, 3-, and 4-dimensional geometric random graphs with the same number of nodes and edges as the PPI network, respectively, while "3x" and "6x" indicate that these model networks have the same number of nodes, but 3 and 6 times as many edges as the PPI network, respectively.

	6 minute			3 minute			30 second			3 second		
Graph	num.	%	D	num.	%	D	num.	%	D	num.	%	D
GEO-2D 1	N/A	N/A	N/A	1	0.04%	0.01	N/A	N/A	N/A	449	18.65%	4.32
GEO-2D 2	0	0%	0	0	0%	0	N/A	N/A	N/A	420	17.42%	4.31
GEO-2D 3	0	0%	0	0	0%	0	N/A	N/A	N/A	424	17.69%	5.79
GEO-2D 4	0	0%	0	0	0%	0	N/A	N/A	N/A	381	15.85%	4.55
GEO-2D 5	0	0%	0	0	0%	0	N/A	N/A	N/A	313	12.97%	3.60
GEO-3D 1	0	0%	0	0	0%	0	1	0.04%	0.02	461	18.79%	5.05
GEO-3D 2	0	0%	0	0	0%	0	0	0%	0	515	20.98%	6.05
GEO-3D 3	0	0%	0	0	0%	0	0	0%	0	513	20.90%	5.30
GEO-3D 4	0	0%	0	0	0%	0	0	0%	0	412	16.78%	4.38
GEO-3D 5	0	0%	0	0	0%	0	0	0%	0	310	12.63%	3.70
GEO-4D 1	0	0%	0	0	0%	0	0	0%	0	483	19.67%	5.52
GEO-4D 2	0	0%	0	0	0%	0	N/A	N/A	N/A	435	17.72%	5.00
GEO-4D 3	0	0%	0	0	0%	0	N/A	N/A	N/A	469	19.10%	5.78
GEO-4D 4	0	0%	0	0	0%	0	N/A	N/A	N/A	507	20.53%	5.33
GEO-4D 5	0	0%	0	0	0%	0	N/A	N/A	N/A	451	18.37%	5.62

Table 9: The number of unprocessed nodes by the TLNP heuristic search algorithm in the geometric random networks corresponding to the yeast top 11000 PPI network and the distances from the exact graphlet counts. Columns denoted by "num." present number of nodes unfinished by the TLNP with 6 minute, 3 minute, 30 second, and 3 second cut-off times, respectively, while columns denoted by "%" present fractions of unprocessed nodes by these experiments with respect to the total number of nodes in the corresponding network. Columns denoted by "D" present the distance of the heuristic graphlet distribution from the exact one. "N/A" indicates that experiments for the given network and time bound were not needed, or were not performed. As before, "GEO-2D", "GEO-3D", and "GEO-4D" denote the 2-, 3-, and 4-dimensional geometric random graphs with the same number of nodes and edges as the PPI network, respectively.

			Degree			
Network G	Radius(G)	Diam(G)	min.	max.	avg.	stDev.
GEO-3D-3x 1	7	13	1	31	14.43	5.16
GEO-3D-3x 2	8	14	2	27	13.74	4.81
GEO-3D-3x 3	8	14	2	28	13.78	4.71
GEO-3D-6x 1	5	10	5	56	28.86	8.96
GEO-3D-6x 2	5	9	4	52	29.36	9.83
GEO-3D-6x 3	5	10	4	58	28.59	9.31
GEO-3D-6x 4	5	9	7	55	28.71	9.66
GEO-3D-6x 5	5	10	4	55	28.95	9.62

Table 10: Network statistics for geometric random networks corresponding to the yeast high-confidence PPI network. Columns denoted by "Radius(G)" and "Diameter(G)" present diameters and radii of these networks, respectively (note that these networks are connected), while other columns give degree statistics: minimum degree, maximum degree, average degree, and the standard deviation of the degrees in these networks, respectively.

		Eccentricity				Degree			
Network	num.	min.	max.	avg.	stDev.	min.	max.	avg.	stDev.
GEO-3D-3x 1, 3s	327	8	13	11.20	1.14	1	17	9.83	3.23
GEO-3D-3x 2, 3s	441	8	14	11.27	1.10	2	17	9.97	3.18
GEO-3D-3x 3, 3s	413	8	14	11.52	1.06	2	17	9.74	3.16
GEO-3D-3x 1, 10s	848	7	13	10.69	1.16	1	22	13.10	4.13
GEO-3D-3x 2, 10s	911	8	14	10.78	1.20	2	23	13.01	4.23
GEO-3D-3x 3, 10s	938	8	14	10.88	1.22	2	23	13.34	4.40
GEO-3D-3x 1, 30s	970	7	13	10.58	1.17	1	29	14.22	4.94
GEO-3D-3x 2, 30s	984	8	14	10.71	1.22	2	27	13.71	4.77
GEO-3D-3x 3, 30s	986	8	14	10.82	1.24	2	26	13.76	4.69

Table 11: Eccentricity and degree statistics for nodes in the GEO-3D networks corresponding to the yeast high-confidence PPI network that were finished by the TLNP experiments within given cut-off times (3, 10, or 30 seconds, denoted by "3s", "10s", and "30s", respectively). The column denoted by "num." gives the number of finished nodes, while "min", "max", "avg." and "stDev" denote minimum, maximum, average, and standard deviation of eccentricities and degrees of these nodes, respectively.

		Eccentricity				Degree			
Network	num.	min.	max.	avg.	stDev.	min.	max.	avg.	stDev.
GEO-3D-3x 1, 3s	661	7	13	10.26	1.05	4	31	16.70	4.36
GEO-3D-3x 2, 3s	547	8	13	10.25	1.12	7	27	16.78	3.59
GEO-3D-3x 3, 3s	575	8	13	10.31	1.11	8	28	16.67	3.30
GEO-3D-3x 1, 10s	139	8	12	9.81	0.91	10	31	22.45	3.20
GEO-3D-3x 2, 10s	77	8	12	9.83	1.15	18	27	22.40	1.67
GEO-3D-3x 3, 10s	50	8	12	9.62	0.95	18	28	21.96	1.84
GEO-3D-3x 1, 30s	18	8	11	9.94	0.87	12	31	25.72	4.04
GEO-3D-3x 2, 30s	4	9	11	10	0.82	12	26	21.75	6.55
GEO-3D-3x 3, 30s	2	9	10	9.5	0.71	18	28	23	7.07

Table 12: Eccentricity and degree statistics for nodes in the GEO-3D networks corresponding to the yeast high-confidence PPI network that remained unfinished by the TLNP within given cut-off times (3, 10, or 30 seconds, denoted by "3s", "10s", and "30s", respectively). The column denoted by "num." gives the number of unfinished nodes, while "min", "max", "avg." and "stDev" denote minimum, maximum, average, and standard deviation of eccentricities and degrees of these nodes, respectively.

		Eccentricity				Degree			
Network	num.	min.	max.	avg.	stDev.	min.	max.	avg.	stDev.
GEO-3D-6x 1, 3s	11	8	10	9	0.45	5	12	8.45	2.88
GEO-3D-6x 2, 3s	7	8	9	8.86	0.38	4	14	9.86	3.72
GEO-3D-6x 3, 3s	2	9	10	9.5	0.71	4	5	4.5	0.71
GEO-3D-6x 4, 3s	3	9	9	9	0	7	9	8.33	1.15
GEO-3D-6x 5, 3s	7	9	10	9.29	0.49	4	10	6.71	2.56
GEO-3D-6x 1, 30s	144	7	10	8.59	0.60	5	26	14.31	4.29
GEO-3D-6x 2, 30s	158	7	9	8.43	0.60	4	24	15.24	4.01
GEO-3D-6x 3, 30s	174	7	10	8.44	0.66	4	26	15.70	4.03
GEO-3D-6x 4, 30s	200	7	9	8.49	0.58	7	27	15.99	3.98
GEO-3D-6x 5, 30s	178	7	10	8.60	0.59	4	25	15.13	4.35
GEO-3D-6x 1, 3m	790	6	10	7.72	0.79	5	41	26.29	7.89
GEO-3D-6x 2, 3m	730	5	9	7.67	0.79	4	41	25.38	7.90
GEO-3D-6x 3, 3m	790	5	10	7.71	0.84	4	41	25.52	7.34
GEO-3D-6x 4, 3m	765	6	9	7.79	0.79	7	41	25.16	7.44
GEO-3D-6x 5, 3m	769	6	10	7.85	0.76	4	42	25.66	7.90

Table 13: Eccentricity and degree statistics for nodes in the GEO-3D networks corresponding to the yeast high-confidence PPI network that were finished by the TLNP with given time bounds (3 seconds, 30 seconds, and 3 minutes, denoted by “3s”, “30s”, and “3m”, respectively). The column denoted by “num.” gives the number of unfinished nodes, while “min”, “max”, “avg.” and “stDev” denote minimum, maximum, average, and standard deviation of eccentricities and degrees of these nodes, respectively.

		Eccentricity				Degree			
Network	num.	min.	max.	avg.	stDev.	min.	max.	avg.	stDev.
GEO-3D-6x 1, 3s	977	5	10	7.53	0.84	6	56	29.09	8.74
GEO-3D-6x 2, 3s	981	5	9	7.41	0.85	7	52	29.50	9.72
GEO-3D-6x 3, 3s	986	5	10	7.50	0.90	6	58	28.64	9.28
GEO-3D-6x 4, 3s	985	5	9	7.55	0.87	7	55	28.78	9.61
GEO-3D-6x 5, 3s	981	5	10	7.60	0.86	5	55	29.11	9.47
GEO-3D-6x 1, 30s	843	5	9	7.37	0.76	13	56	31.34	6.98
GEO-3D-6x 2, 30s	830	5	9	7.23	0.76	9	52	32.05	8.17
GEO-3D-6x 3, 30s	814	5	9	7.30	0.83	15	58	31.35	7.70
GEO-3D-6x 4, 30s	788	5	9	7.32	0.76	15	55	31.94	7.85
GEO-3D-6x 5, 30s	810	5	9	7.40	0.77	16	55	31.99	7.58
GEO-3D-6x 1, 3m	198	5	9	6.87	0.75	19	56	39.11	4.58
GEO-3D-6x 2, 3m	258	5	8	6.71	0.59	24	52	40.65	4.61
GEO-3D-6x 3, 3m	198	5	9	6.70	0.69	18	58	40.86	5.66
GEO-3D-6x 4, 3m	223	5	9	6.77	0.62	21	55	40.90	5.67
GEO-3D-6x 5, 3m	219	5	9	6.79	0.71	11	55	40.52	5.12

Table 14: Eccentricity and degree statistics for nodes in the GEO-3D networks corresponding to the yeast high-confidence PPI network that remained unfinished by the TLNP with given time bounds (3 seconds, 30 seconds, and 3 minutes, denoted by “3s”, “30s”, and “3m”, respectively). The column denoted by “num.” gives the number of unfinished nodes, while “min”, “max”, “avg.” and “stDev” denote minimum, maximum, average, and standard deviation of eccentricities and degrees of these nodes, respectively.

	Largest Connected Component		Degree			
Net. G	Radius	Diameter	min.	max.	avg.	stDev.
GEO-2D 1	27	52	0	19	9.09	3.01
GEO-2D 2	27	53	0	20	9.09	2.94
GEO-2D 3	26	52	0	22	9.15	3.19
GEO-2D 4	27	53	0	19	9.09	2.96
GEO-2D 5	27	52	0	18	9.04	2.85
GEO-3D 1	13	24	0	21	9.00	3.22
GEO-3D 2	13	24	0	20	8.98	3.30
GEO-3D 3	13	25	1	19	9.02	3.23
GEO-3D 4	13	25	0	19	8.99	3.28
GEO-3D 5	13	25	1	21	8.97	3.26
GEO-4D 1	10	17	0	23	8.92	3.84
GEO-4D 2	10	17	0	21	8.89	3.84
GEO-4D 3	10	18	0	24	8.93	3.95
GEO-4D 4	10	18	0	22	9.03	3.89
GEO-4D 5	10	18	0	24	8.95	3.82

Table 15: Network statistics for geometric random networks corresponding to the yeast top 11000 PPI network. The columns denoted by “Radius” and “Diameter” present radii and diameters of the largest connected components of these networks, while other columns give degree statistics of the entire network: minimum degree, maximum degree, average degree, and the standard deviation of the degrees, respectively.

		Eccentricity				Degree			
Network	num.	min.	max.	avg.	stDev.	min.	max.	avg.	stDev.
GEO-2D 1, 3s	449	27	49	39.95	4.52	8	19	13.00	1.98
GEO-2D 2, 3s	420	27	51	39.20	5.50	4	20	13.05	2.06
GEO-2D 3, 3s	424	26	51	39.76	6.24	9	22	13.83	2.13
GEO-2D 4, 3s	381	28	49	37.48	4.85	9	19	13.55	1.81
GEO-2D 5, 3s	313	27	48	38.41	5.45	9	18	13.47	1.51
GEO-3D 1, 3s	461	14	22	18.05	1.64	6	21	13.51	1.82
GEO-3D 2, 3s	515	14	23	18.57	2.00	9	20	13.61	1.70
GEO-3D 3, 3s	513	14	24	17.75	1.79	10	19	13.48	1.58
GEO-3D 4, 3s	412	13	22	17.29	1.66	10	19	13.86	1.57
GEO-3D 5, 3s	310	15	24	18.46	1.45	6	21	14.21	2.00
GEO-4D 1, 3s	483	10	15	12.29	1.10	1	23	14.09	2.97
GEO-4D 2, 3s	435	10	15	12.36	0.95	11	21	14.64	2.01
GEO-4D 3, 3s	469	10	17	12.62	1.08	3	24	14.57	2.61
GEO-4D 4, 3s	507	10	15	12.54	0.85	10	22	14.44	2.06
GEO-4D 5, 3s	451	10	16	12.92	1.33	9	24	14.57	2.13

Table 16: Eccentricity and degree statistics for nodes in the GEO networks corresponding to the yeast top 11000 PPI network that were unfinished by TLNP with the given 3 second time bound (denoted by "3s"). The column denoted by "num." gives the number of unfinished nodes, while "min", "max", "avg." and "stDev" denote minimum, maximum, average, and standard deviation of eccentricities and degrees of these nodes, respectively.

	3 second					targeted 1%			
Graph	Num.	%	D	Finished Time	Total Time	Num.	%	D	Total Time
GEO-3D-6x 1	977	98.89%	40.60	20.12s	49m 11.12s	979	99.09%	33.14	3m 46.75s
GEO-3D-6x 2	981	99.29%	32.76	15.28s	49m 18.28s	979	99.09%	50.66	3m 12.41s
GEO-3D-6x 3	986	99.80%	67.22	3.43s	49m 21.43s	979	99.09%	47.68	2m 55.88s
GEO-3D-6x 4	985	99.70%	39.02	6.34s	49m 21.34s	979	99.09%	42.85	3m 0.37s
GEO-3D-6x 5	981	99.29%	35.99	13.42s	49m 16.42s	979	99.09%	32.10	2m 5.39s

Table 17: The number of unprocessed nodes by the TLNP heuristic with 3 second time cut-off and TN search with full processing of the targeted 1% of the nodes, respectively, in 3-dimensional geometric random networks with the same number of nodes and six times as many edges as the yeast high-confidence PPI network. Columns denoted by "Num." and "%" present the number and percentage (out of the total number of nodes in the network) of unprocessed nodes, respectively, while columns denoted by "D" give distance between the estimated and the exhaustive graphlet frequency distributions. "Finished Time" is the sum of times TLNP took to process the nodes that got finished within 3 seconds, while the "Total Time" in the "3 second" column is the sum of 3 seconds per unfinished node plus the time it took to complete the finished nodes (for example, in the first row, 49m 11.12s = 977 $\times$ 3s + 20.12s). "Total Time" in the "targeted 1%" column represents the total time that TNP took to process the selected nodes.

	3 second					targeted 2%			
Graph	Num.	%	D	Finished Time	Total Time	Num.	%	D	Total Time
GEO-3D-6x 1	977	98.89%	40.60	20.12s	49m 11.12s	969	98.08%	35.63	7m 29.47s
GEO-3D-6x 2	981	99.29%	32.76	15.28s	49m 18.28s	969	98.08%	40.43	7m 43.69s
GEO-3D-6x 3	986	99.80%	67.22	3.43s	49m 21.43s	969	98.08%	35.92	6m 10.64s
GEO-3D-6x 4	985	99.70%	39.02	6.34s	49m 21.34s	969	98.08%	38.72	5m 27.11s
GEO-3D-6x 5	981	99.29%	35.99	13.42s	49m 16.42s	969	98.08%	38.80	6m 23.84s

Table 18: The number of unprocessed nodes by the TLNP heuristic with 3 second cut-off time and TNP heuristic with full processing of the targeted 2%, respectively, of the graph nodes in the 3-dimensional geometric random networks with the same number of nodes and six times as many edges as in the yeast high-confidence PPI network. Columns denoted by "Num." and "%" present the number and percentage (out of the total number of nodes in the network) of unprocessed nodes, respectively, while columns denoted by "D" give the distance between the heuristic and the exhaustive graphlet frequency distributions. "Finished Time" is the sum of the times it took to process the nodes that got finished within 3 seconds, while the "Total Time" in the "3 second" column is the sum of 3 seconds per unfinished node plus the time it took to complete the finished nodes (for example, in the first row, 49m 11.12s = 977 $\times$ 3s + 20.12s). "Total Time" in the "targeted 2%" column represents the total time that TNP took to process the selected nodes.

	Num. Unfinished	% Unfinished	Distance
YHC SF 1	6	0.6%	22.64
YHC SF 2	8	0.8%	44.23
YHC SF 3	6	0.6%	28.45
YHC SF 4	4	0.4%	27.04
YHC SF 5	5	0.5%	28.46
Y11K SF 1	80	3.3%	32.14
FH SF 1	8	0.2%	54.82
FE SF 1	132	1.9%	57.84
W SF 1	15	0.5%	72.01
YHC ER-DD 1	6	0.6%	5.34
YHC ER-DD 2	7	0.7%	7.20
YHC ER-DD 3	6	0.6%	5.85
YHC ER-DD 4	8	0.8%	7.59
YHC ER-DD 5	7	0.7%	8.09
Y11K ER-DD 1	232	9.7%	27.34
Y11K ER-DD 2	233	9.7%	26.89
Y11K ER-DD 3	240	10%	27.06
Y11K ER-DD 4	240	10%	26.70
Y11K ER-DD 5	238	9.9%	27.06
FH ER-DD 1	1	0.02%	11.71
FE ER-DD 1	232	3.3%	33.59
W ER-DD 1	15	0.5%	66.33

Table 19: Graphlet frequency distances between the results of the exhaustive and the TLNP with 3 minute cut-off time graphlet searches for SF and ER-DD networks corresponding to different PPI networks. “YHC SF” denotes scale-free networks corresponding to the yeast high confidence PPI network. Similarly, “Y11K SF”, “FH SF”, “FE SF”, and “W SF” denote scale-free networks corresponding to the yeast top 11000, high-confidence fruitfly, entire fruitfly, and worm PPI networks, respectively, while “Y11K ER-DD”, “FH ER-DD”, “FE ER-DD”, and “W ER-DD” denote ER-DD networks corresponding to the yeast top 11000, high-confidence fruitfly, entire fruitfly, and worm PPI networks, respectively.

			Degree			
Net. G	Radius(G)	Diameter(G)	min.	max.	avg.	stDev.
YHC SF 1	4	7	2	64	4.98	5.89
YHC SF 2	4	7	2	138	4.96	7.04
YHC SF 3	4	7	2	92	4.98	6.56
YHC SF 4	4	7	2	98	4.97	6.37
YHC SF 5	4	6	2	103	4.97	6.71
Y11K SF 1	3	5	4	176	9.15	10.79
FH SF 1	12	24	1	100	2.00	3.22
FE SF 1	4	7	3	321	6.00	8.62
W SF 1	5	10	1	205	3.44	5.84
YHC ER-DD 1	4	6	1	51	4.97	7.45
YHC ER-DD 2	4	6	1	51	4.97	7.45
YHC ER-DD 3	4	6	1	51	4.97	7.45
YHC ER-DD 4	4	6	1	51	4.97	7.45
YHC ER-DD 5	4	6	1	51	4.97	7.45
Y11K ER-DD 1	4	5	1	114	9.16	15.52
Y11K ER-DD 2	4	5	1	114	9.16	15.52
Y11K ER-DD 3	4	5	1	114	9.16	15.52
Y11K ER-DD 4	4	5	1	114	9.16	15.52
Y11K ER-DD 5	4	5	1	114	9.16	15.52
FH ER-DD 1	18	33	1	38	2.02	1.81
FE ER-DD 1	5	7	1	173	5.73	9.12
W ER-DD 1	6	9	1	187	3.44	7.10

Table 20: Network statistics for the SF and ER-DD networks corresponding to the PPI networks. The columns denoted by "Radius(G)" and "Diameter(G)" present radii and diameters of these networks, respectively (these networks, except FH ER-DD 1, are connected; for FH ER-DD 1, the radius and the diameter of the largest connected component are reported), while other columns give degree statistics of the entire network: minimum degree, maximum degree, average degree, and the standard deviation of the degrees, respectively. As before, "YHC SF" denotes scale-free networks corresponding to the yeast high confidence PPI network etc. (see caption of Supplementary Table 19).

		Eccentricity				Degree			
Network	num.	min.	max.	avg.	stDev.	min.	max.	avg.	stDev.
YHC SF 1	6	4	5	4.17	0.41	43	64	59	7.97
YHC SF 2	8	4	5	4.25	0.46	40	138	63.12	32.75
YHC SF 3	6	4	5	4.17	0.41	60	92	70.33	11.27
YHC SF 4	4	4	4	4	0	67	98	78	13.74
YHC SF 5	5	4	4	4	0	48	103	74.4	19.65
Y11K SF 1	80	3	4	3.99	0.11	24	176	54.25	29.80
FH SF 1	8	12	15	13.62	1.06	33	100	52.87	20.78
FE SF 1	132	4	6	4.96	0.23	20	321	47.27	38.84
W SF 1	15	5	7	6.13	0.52	36	205	61.33	44.19
YHC ER-DD 1	6	4	5	4.17	0.41	39	51	47.5	4.37
YHC ER-DD 2	7	4	5	4.14	0.38	46	51	48.29	1.98
YHC ER-DD 3	6	4	5	4.33	0.52	46	51	48.67	1.86
YHC ER-DD 4	8	4	5	4.12	0.35	39	51	46.62	4.17
YHC ER-DD 5	7	4	4	4	0	39	51	47	4.20
Y11K ER-DD 1	232	4	4	4	0	21	114	47.38	24.58
Y11K ER-DD 2	233	4	4	4	0	15	114	47.24	24.63
Y11K ER-DD 3	240	4	4	4	0	16	114	46.68	24.46
Y11K ER-DD 4	240	4	4	4	0	15	114	46.5	24.64
Y11K ER-DD 5	238	4	4	4	0	15	114	46.76	24.59
FH ER-DD 1	1	21	21	21	N/A	38	38	38	N/A
FE ER-DD 1	232	5	5	5	0	23	173	41.84	19.14
W ER-DD 1	24	6	7	6.08	0.28	25	187	62.92	34.75

Table 21: Eccentricity and degree statistics for nodes in the SF and ER-DD networks corresponding to the PPI networks that were unfinished by TLNP with 3 minute cut-off time. The column denoted by "num." gives the number of unfinished nodes, while "min", "max", "avg." and "stDev" denote minimum, maximum, average, and standard deviation of eccentricities and degrees of these nodes, respectively. As before, "YHC SF" denotes scale-free networks corresponding to the yeast high confidence PPI network etc. (see caption of Supplementary Table 19).

PPI Net	Cut-off	Num. Unf.	% Unf.	Dist.	CPU Time Heur.	CPU Time Exhaust.
YHC	180 sec.	14	1.4%	7.58	47m 49.29s	8h 57m 16.63s
YHC	30 sec.	36	3.6%	15.14	17m 45.32s	8h 57m 16.63s
YHC	3 sec.	114	11.5%	37.94	3m 26.50s	8h 57m 16.63s
Y11K	180 sec.	250	10.4%	37.86	6h 39m 28.98s	9d 14h 3m 46.08s
Y11K	30 sec.	430	17.9%	42.89	1h 42m 15.42s	9d 14h 3m 46.08s
Y11K	3 sec.	920	38.3%	68.44	16m 11.70s	9d 14h 3m 46.08s
FH	180 sec.	1	0.02%	9.95	13m 21.88s	9m 36.17s
FH	30 sec.	2	0.04%	11.14	13m 45.90s	9m 36.17s
FH	3 sec.	24	0.5%	17.90	10m 26.44s	9m 36.17s
FE	180 sec.	331	4.7%	22.35	20h 25m 26.35s	8d 7h 28m 33.16s
FE	30 sec.	1104	15.8%	33.19	7h 26m 1.23s	8d 7h 28m 33.16s
FE	3 sec.	3961	56.7%	83.69	1h 4m 20.61s	8d 7h 28m 33.16s
W	180 sec.	93	3.0%	15.41	1h 44m 35.65s	$>$ 5.3 months
W	30 sec.	180	5.8%	21.51	55m 28.53s	$>$ 5.3 months
W	3 sec.	388	12.5%	28.08	27m 12.74s	$>$ 5.3 months

Table 22: CPU time in seconds (s), minutes (m), hours (h), days (d), or months taken by the TLNP experiments with different cut-off times and the exhaustive searches for the five PPI networks. The "Dist." column presents distances between the exhaustive and the heuristic graphlet counts except for the worm ("W") PPI network where the distances are between the 4 hour time limited TLNP graphlet count and TLNP graphlet counts with lower cut-off times (this is because we could not process this PPI network exhaustively). The PPI networks are denoted as before (see the caption of Supplementary Table 5).

PPI Net	Cut-off	Num. Unf.	% Unf.	Dist.	CPU Time Heur.	CPU Time Exhaust.
GEO-3D-6x 1	180 sec.	198	20.04%	2.11	14h 57m 12.66s	18h 52m
GEO-3D-6x 1	30 sec.	843	88.09%	16.29	34m 37.66s	18h 52m
GEO-3D-6x 1	3 sec.	977	98.89%	40.60	20.12s	18h 52m
GEO-3D-6x 2	180 sec.	258	26.11%	3.43	13h 40m 26.42s	21h 38m
GEO-3D-6x 2	30 sec.	830	84.00%	17.84	43m 16.6s	21h 38m
GEO-3D-6x 2	3 sec.	981	99.29%	32.76	15.28s	21h 38m
GEO-3D-6x 3	180 sec.	198	20.04%	2.90	13h 40m 17.67s	19h 18m
GEO-3D-6x 3	30 sec.	814	82.39%	14.94	48m 18.6s	19h 18m
GEO-3D-6x 3	3 sec.	986	99.80%	67.22	3.43s	19h 18m
GEO-3D-6x 4	180 sec.	223	23.58%	3.52	12h 59m 25.53s	18h 47m
GEO-3D-6x 4	30 sec.	788	79.76%	14.80	52m 36.6s	18h 47m
GEO-3D-6x 4	3 sec.	985	99.70%	39.02	6.34s	18h 47m
GEO-3D-6x 5	180 sec.	219	22.16%	2.72	14h 7m 36.24s	18h 24m
GEO-3D-6x 5	30 sec.	810	81.98%	17.48	47m 4.13s	18h 24m
GEO-3D-6x 5	3 sec.	981	99.29%	35.99	13.42s	18h 24m

Table 23: CPU times in seconds (s), minutes (m), or hours (h) taken by the TLNP experiments with different cut-off times and the exhaustive search algorithms for geometric random networks corresponding to the yeast high-confidence PPI network. "Num. Unf." and "% Unf." present the number and percentage of unfinished nodes with the specified cut-off time. The "Dist." column presents distances between the exhaustive and the heuristic graphlet distributions.

PPI Net	Cut-off	Num. Unf.	% Unf.	Dist.	CPU Time Heur.	CPU Time Exhaust.
SF 1	3 min.	6	0.6%	22.64	20m 12.68s	1h 52m 39s
SF 2	3 min.	8	0.8%	44.23	18m 42.99s	6h 06m 06s
SF 3	3 min.	6	0.6%	28.45	19m 40.53s	2h 59m 12s
SF 4	3 min.	4	0.4%	27.04	24m 34.03s	3h 08m 02s
SF 5	3 min.	5	0.5%	28.46	26m 23.95s	6h 36m 50s

Table 24: CPU time in seconds (s), minutes (m), or hours (h) taken by the TLNP with 3 minute node processing time limit and the exhaustive search algorithm for SF networks corresponding to the yeast high-confidence PPI network. "Num. Unf." and "% Unf." present the number and percentage of unfinished nodes with the specified cut-off time. The "Dist." column presents distances between the exhaustive and the heuristic graphlet counts.

PPI Net	Cut-off	Num. Unf.	% Unf.	Dist.	CPU Time Heur.	CPU Time Exhaust.
ER-DD 1	3 min.	6	0.6%	5.34	46m 53.54s	58m 43s
ER-DD 2	3 min.	7	0.7%	7.20	43m 04.37s	58m 13s
ER-DD 3	3 min.	6	0.6%	5.85	47m 44.47s	58m 38s
ER-DD 4	3 min.	8	0.8%	7.59	40m 43.17s	59m 42s
ER-DD 5	3 min.	7	0.7%	8.09	43m 02.38s	59m 12s

Table 25: CPU times in seconds (s), minutes (m), or hours (h) taken by the TLNP with 3 minute node processing time limit and the exhaustive search algorithm for ER-DD networks corresponding to the yeast high-confidence PPI network. "Num. Unf." and "% Unf." present the number and percentage of unfinished nodes with the specified cut-off time. The "Dist." column presents distances between the exhaustive and the heuristic graphlet counts.

PPI Net	Cut-off	Num. Fin.	% Fin.	CPU Time Heur.	D1	D2
GEO-3D 1	90 sec	4289	4.29%	2d 16h 24m 28.34s	23.74	26.11
GEO-3D 1	60 sec	1504	1.50%	16h 44m 23.27s	29.27	31.24
GEO-3D 1 rewire 10% 1	90 sec	1957	1.96%	1d 13h 35m 31.23s	31.67	31.38
GEO-3D 1 rewire 10% 2	90 sec	2147	2.15%	1d 14h 24m 01.32s	32.86	32.67
GEO-3D 1 rewire 10% 3	90 sec	2324	2.32%	1d 12h 53m 44.95s	36.84	36.41
GEO-3D 1 rewire 20% 1	120 sec	2859	2.86%	2d 11h 38m 50.36s	46.55	44.76
GEO-3D 1 rewire 20% 2	120 sec	2523	2.52%	2d 04h 29m 10.07s	47.41	45.61
GEO-3D 1 rewire 20% 3	120 sec	2513	2.51%	2d 04h 54m 38.34s	47.12	45.32
GEO-3D 1 rewire 30% 1	180 sec	6089	6.10%	7d 13h 20m 14.61s	54.64	52.69
GEO-3D 1 rewire 30% 2	180 sec	5949	5.95%	7d 08h 55m 46.35s	55.05	53.11
GEO-3D 1 rewire 30% 3	180 sec	5398	5.40%	6d 16h 38m 14.27s	54.79	52.84

Table 26: CPU time in seconds (s), minutes (m), hours (h), or days (d) taken by TLNP with different cut-off times for a geometric random network with 100,000 nodes and 750,000 edges, as well as for networks obtained by rewiring edges in this network. "D1" and "D2" are the distances of the estimated graphlet distribution of this large GEO-3D network from the exact graphlet counts of a GEO-3D network corresponding to the yeast high-confidence PPI network with three and six times as many edges, respectively.

Network	$2\|V\|$	$\|V\|$	$\|V\|/2$	$\|V\|/4$	$\|V\|/8$
PPI	47.35	47.71	48.17	48.16	49.23
ER	83.50	86.86	95.05	97.79	110.06
ER-DD	no runs	99.89	104.05	102.00	109.96
SF	80.80	83.54	88.00	91.87	91.77
GEO-2D	28.37	29.32	29.01	31.40	32.90
GEO-3D	33.71	36.97	34.48	34.84	34.84
GEO-4D	33.10	35.00	33.74	34.63	36.95
GEO-3Dx3	46.73	44.02	44.42	46.46	48.52
GEO-3Dx6	50.10	51.45	51.36	50.00	53.27

Table 27: Graphlet frequency distances between the average graphlet frequencies obtained by the NLS experiments with parameters 2, 5, 3, 1 and the results of the exhaustive search algorithm for the yeast high confidence PPI network and its model networks. The averages of 10 experiments for each of the seed node numbers ($2|V|, |V|, |V|/2, |V|/4$, and $|V|/8$) were taken.

Network	$8\|V\|$	$4\|V\|$	$2\|V\|$	$\|V\|$	$\|V\|/2$	$\|V\|/4$	$\|V\|/8$
PPI	46.41	46.29	46.08	46.40	46.64	46.53	46.46
ER	no runs	no runs	no runs	107.73	114.64	120.95	121.39
ER-DD	113.61	116.76	120.42	121.42	128.65	131.72	138.84
SF	no runs	no runs	107.92	115.44	115.70	119.82	123.73
GEO-2D	37.42	37.18	36.07	34.99	36.00	38.59	38.36
GEO-3D	40.43	39.37	39.79	39.93	41.54	45.62	41.67
GEO-4D	47.69	48.06	48.14	47.89	50.10	48.79	51.61
GEO-3Dx3	48.18	49.08	49.10	48.58	48.44	48.98	52.32
GEO-3Dx6	56.02	56.49	55.62	54.69	56.73	56.40	59.44

Table 28: Graphlet frequency distances between the results of the heuristic NLS algorithm with parameters 20, 20, 10, 2 and different seed node set sizes ($8|V|, 4|V|, 2|V|, |V|, |V|/2, |V|/4$, and $|V|/8$) and the exact graphlet frequencies for the yeast high confidence PPI network and the corresponding model networks. Distances are computed between the exact graphlet frequencies and the average of the estimated graphlet frequencies for each seed node set size (also see caption of Supplementary Table 30). "no runs" indicates that no experiments were done for that particular graph and the selected number of seed nodes, since they would take too much CPU time.

Network	$\|V\|/8$	$\|V\|/4$	$\|V\|/2$	$\|V\|$	$2\|V\|$
PPI	9m 1.35s	15m 25.03s	26m 44.69s	1h 1m 43.03s	1h 34m 3.01s
ER	1d 12h 55m 35s	4d 1h 55m 46s	9d 4h 59m 0s	11d 18h 58m 40s	no runs
ER-DD	38m 29.87s	1h 37m 41.43s	3h 31m 26.2s	5h 22m 41.2s	13h 25m 56.9s
SF	16h 6m 40.5s	1d 6h 2m 13s	1d 23h 23m 46s	5d 23h 7m 44s	8d 13h 18m 20s
GEO-2D	12.06s	27.40s	38.45s	1m 44.27s	2m 44.33s
GEO-3D	20.93s	57.86s	2m 46.64s	4m 50.64s	6m 7.623s
GEO-4D	1m 41.17s	4m 18.39s	5m 47.25s	16m 30.466s	23m 31.25s
GEO-3Dx3	3m 27.77s	6m 53.41s	8m 31.15s	23m 47.81s	40m 36.43s
GEO-3Dx6	3m 11.95s	6m 37.32s	15m 41.14s	40m 1.21s	51m 22.3s

Table 29: CPU times taken by the NLS experiments with parameters 2, 5, 3, 1 and different seed node set sizes ($|V|/8, |V|/4, |V|/2, |V|$, and $2|V|$) for yeast high confidence PPI network and the corresponding model networks. The mean is given for 10 runs in most of the seed size categories (the exceptions are: in the $|V|/8$ category, the average is taken of 7 runs for the ER network; in the $|V|/4$ category, the average is taken of 8 runs for the ER and 9 runs for the SF network; in the $|V|/2$ category, the average is taken of 8 runs for the SF and 9 runs for the ER network; in the $|V|$ category, the average is taken of 3 runs of the ER network; in the $2|V|$ category, the average is taken of 7 runs for the SF and 8 runs for the ER-DD network; in the $2|V|$ category, the average is taken of 2 runs for the GEO-3Dx6 network). "no runs" indicates that no experiments were done for that particular graph and the selected number of seed nodes, since they would take too much CPU time.

Network	$\|V\|/8$	$\|V\|/4$	$\|V\|/2$	$\|V\|$	$2\|V\|$	$4\|V\|$	$8\|V\|$
PPI	5m 40.20s	12m 0.98s	24m 43.79s	47m 26.62s	1h 44m 48.06s	3h 39m 54.1s	8h 58m 1.4s
ER	1d 18h 7m 2s	3d 9h 1m 23s	7d 8h 29m 31s	12d 22h 18m 0s	no runs	no runs	no runs
ER-DD	52m 42.45s	1h 13m 10.92s	4h 15m 35.1s	6h 17m 29.2s	10h 22m 10.9s	21h 20m 58.9s	2d 3h 14m 36s
SF	16h 19m 28.6s	1d 1h 2m 22s	2d 8h 2m 5s	4d 8h 40m 29s	8d 5h 31m 53s	no runs	no runs
GEO-2D	29.90s	32.69s	53.17s	2m 3.89s	3m 40.45s	7m 51.28s	17m 59.27s
GEO-3D	26.063s	58.589s	2m 35.99s	3m 43.09s	9m 41.52s	14m 47.11s	34m 26.17s
GEO-4D	1m 33.30s	3m 22.39s	11m 30.939s	13m 12.507s	25m 51.97s	1h 2m 48.67s	2h 1m 41.39s
GEO-3Dx3	4m 45.19s	6m 12.186s	11m 49.652s	18m 50.66s	34m 54.11s	1h 15m 34.39s	2h 36m 54.14s
GEO-3Dx6	3m 26.07s	7m 28.82s	16m 9.05s	29m 24.35s	1h 6m 43.28s	2h 6m 41.89s	5h 13m 56s

Table 30: CPU times taken by the NLS experiments with parameters 20, 20, 10, 2 and different seed node set sizes ($|V|/8, |V|/4, |V|/2, |V|, 2|V|, 4|V|$, and $8|V|$) for yeast high confidence PPI network and the corresponding model networks. The mean is given for 10 runs in each of the seed size categories (the exceptions are: in the $|V|/2$ category, the average is taken of 9 runs for the SF, and 7 runs for the ER network; in the $|V|$ category, the average is is taken of 5 runs for the SF, 3 runs for the ER, and 7 runs for the ER-DD network; in the $2|V|$ category, the average is taken for 3 runs of the GEO-3D and 7 runs for the SF network; in the $2|V|$ category, the average is of 5 runs for the ER-DD network). "no runs" indicates that no experiments were done for that particular graph and the selected number of seed nodes, since they would take too much CPU time.

	TNP				NLS				
Network G	Speedup	Dist.	Num.	% Nodes	Speedup	Dist.	Num.	% Nodes	Param's
YHC	690	45.91	197	20%	95	46.46	123 = $\frac{\|V\|}{8}$	12%	20, 20, 10, 2
YHC	147	39.37	296	30%	60	49.23	247 = $\frac{\|V\|}{4}$	25%	2, 5, 3, 1
GEO-3D x 6	323	33.14	9	1%	377	53.27	123 = $\frac{\|V\|}{8}$	12%	2, 5, 3, 1
GEO-3D x 6	162	35.63	19	2%	174	50.00	247 = $\frac{\|V\|}{4}$	25%	2, 5, 3, 1

Table 31: Comparisons of TNP and NLS performance for the yeast high-confidence (YHC) PPI and the corresponding geometric random network (GEO-3D x 6) in terms of running time $r = \frac{T_E}{T_H}$ ratio, distance ("Dist."), number of processed (in TNP) or seed (in NLS) nodes ("Num."), percentage of processed (in TNP) or seed (in NLS) nodes ("% Nodes"), and search parameters ("Param's") in NLS heuristic.

References

Giot, L., J. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. Hao, C. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. Stanyon, R. J. Finley, K. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. Shimkets, M. McKenna, J. Chant, and J. Rothberg (2003). A protein interaction map of drosophila melanogaster. *Science 302*(5651), 1727–1736.

Kashtan, N., S. Itzkovitz, R. Milo, and U. Alon (2004). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics 20*, 1746–1758.

Li, S., C. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.-O. Vidalain, J.-D. Han, A. Chesneau, T. Hao, N. Goldberg, DS Li, M. Martinez, J.-F. Rual, P. Lamesch, L. Xu, M. Tewari, S. Wong, L. Zhang, G. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. Gabel, A. Elewa, B. Baumgartner, D. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. Mango, W. Saxton, S. Strome, S. van den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. Gunsalus, J. Harper, M. Cusick, F. Roth, D. Hill, and M. Vidal (2004). A map of the interactome network of the metazoan c. elegans. *Science 303*, 540–543.

Mathon, R. (2004). personal communication.

Mehlhorn, K. and S. Naher (1999). *Leda: A platform for combinatorial and geometric computing*. Cambridge University Press.

Pržulj, N. (2005). *Analyzing Large Biological Networks: Protein-Protein Interactions Example*. Ph. D. thesis, University of Toronto, Canada.

Pržulj, N., D. G. Corneil, and I. Jurisica (2004). Modeling interactome: Scale-free or geometric? *Bioinformatics 20*(18), 3508–3515.

von Mering, C., R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature 417*(6887), 399–403.