Pacific Symposium on Biocomputing 14:178-189 (2009)

DENSE GRAPHLET STATISTICS OF PROTEIN INTERACTION AND RANDOM NETWORKS

R. COLAK*, F. HORMOZDIARI*, F. MOSER *, A. SCHÖNHUTH*,
J. HOLMAN, M. ESTER AND S.C. SAHINALP+

School of Computing Science, Simon Fraser University

* *Joint first authors.* + *Contact author: cenk@cs.sfu.ca*

Understanding evolutionary dynamics from a systemic point of view crucially depends on knowledge about how evolution affects size and structure of the organisms' functional building blocks (*modules*). It has been recently reported that statistics over sparse PPI graphlets can robustly monitor such evolutionary changes. However, there is abundant evidence that in PPI networks modules can be identified with highly interconnected (*dense*) and/or bipartite subgraphs. We count such dense graphlets in PPI networks by employing recently developed search strategies that render related inference problems tractable. We demonstrate that corresponding counting statistics differ significantly between prokaryotes and eukaryotes as well as between ”real” PPI networks and scale free network emulators. We also prove that another class of emulators, the low-dimensional geometric random graphs (GRGs) cannot contain a specific type of motifs, complete bipartite graphs, which are abundant in PPI networks.

1. Introduction

On the biochemical level, life can be explained by gene products facilitating and controlling essential cellular mechanisms, thereby establishing overall cellular viability. To suitably model the structural features that underlie the complex wirings of gene products is presently at the core of computational systems biology. As further explained by the modularity paradigm, global interplay of cellular mechanisms can be decomposed into interaction of functional subunits (*functional modules*), which consist of subsets of gene products that facilitate essential functionalities by concerted actions. Therefore, beyond studying global properties of biomolecular networks (BNs) such as the degree distribution, there has been considerable interest in identifying and quantifying *local* topological properties of biomolecular networks. In particular, small subgraphs or (*graphlets*) that appear significantly more frequently in biomolecular networks than in random graphs can yield insights on cellular functionalities[10]. In order to identify dense graphlets, a variety of methods that count induced subgraphs of different types and node sizes of up to 5, 6 and 7 have been suggested[12,7,5]. In the most recent approach, an algorithm has been developed that counts all non-induced subtrees (as well as certain other ”sparse” graphlets) of a PPI network[1]. Corresponding subtree statistics turned out to be robust similarity measures between PPI networks, hence, in principle, can quantify organismic changes due to *evolutionary dynamics*.

It has been argued that graphlet distributions resulting from graphlet counting algorithms can also be used to assess the suitability of random graph models for emulating the growth of and achieving the observable topological features of BNs. A thorough assessment would be highly desirable, as available models, although correctly accounting for many of the global features of BNs, can be different in terms of local features. For example, while exhibiting degree distributions similar to those of prior models (e.g. the *preferential attachment model* (PAM) which gives rise to scale-free networks), the *generalized duplication model* (GDM) is superior in terms of graphlet distributions that have recently been suggested[7]. This may be explained by that the GDM is the only model under actual consideration that is based on emulation of processes that guide BN growth, i.e. gene duplications. Apart from the GDM and the PAM, the geometric random graph model (GRGM) has recently been introduced as an intuitive alternative option. Although not being biologically motivated, PPI networks were successfully fit to GRGMs of low dimensions[6] (for definitions see Section 3).

1.1. *Motivation: Dense Biomolecular Graphlets*

It has been widely established in the biological literature[10], modules in PPI networks are most likely encoded as highly interconnected (*dense*) regions, i.e., connected subgraphs that contain relatively large numbers of edges. It is thus of interest how these regions change over evolutionary time and whether they play a significant role in how the PPI networks evolve. As a result, determining the number of specific dense graphlets and analyzing how they vary with respect to the ”complexity” of their respective organism is a problem of significant interest. Unfortunately, counting dense graphlets, even approximately, is a highly non-trivial problem; for general graphs it is known to be intractable. In fact, the simpler problem of determining whether an input graph G includes one or more cliques of size k, or any graphlet with density $1 - O(1/n)$, is NP-hard. Furthermore, the number of any specific dense graphlet M in a dense graph G would be exponential with the size of M.

Fortunately it is well known that PPI networks are typically very sparse, with average degree of 7 or less. For such networks, a novel data mining tool for determining whether a given network G includes a dense graphlet M has recently been described[4]. This tool is based on a pruning technique derived from combinatorial observations on dense subgraphs[4]. Although, in the worst case, the running time of this pruning technique is exponential with the size of the dense graphlet it is looking for, it is of considerable interest whether it can be generalized to count a given dense graphlet in a PPI network of interest. Once such statistics are obtained for all dense graphlets one can consider a number of interesting questions.

(i) Are prominent PPI network generation models, in particular GRGM (i-a) and GDM (i-b) in accordance with statistics over such dense graphlets? If not, it is highly unlikely that they sufficiently account for central cellular functionalities.

(ii) Can we use dense graphlet distributions of various organisms, in order to quantify changes in organism complexity due to evolutionary dynamics? In general, it would be interesting to find out whether evolutionary trends towards more complex organisms can be monitored this way. If one accepts that PPI networks have evolved in a duplication oriented procedure, then one might be able to suggest that more evolutionary complex species might have more complex graphlets.

As a last point, we noticed that in networks generated by low-dimensional GRGMs, complete bipartite graphs of the types $K_{n,m}$ where $n, m \geq 3$ do not/can not exist. However, there is evidence that induced bipartite graphs abundantly occur in the PPI networks of E.coli[3] and others. Moreover, there is evidence that complete bipartite graphs in PPI networks are related to "parallel" functional modules which increase cellular flexibility and robustness.Furthermore, complete bipartite graphs of four nodes (known as **bi-fan**) are the main building blocks of dense overlap regulons (a regulon is the set of genes regulated by given transcription factor) [2,9]. Bi-fan generalizations to larger patterns with row of inputs and row of outputs (bipartite graphs of larger size) are also abundant in PPI networks [2]. Therefore, we put particular emphasis on bipartite graphlet statistics.

1.2. *Contributions*

Counting graphlets (induced or non-induced form) with more than 7 nodes has been a challenging computational task. Recent advances allow improvement for sparse, non-induced graphlets: it is now possible to count trees and bounded treewidth graphlets (with very small treewidth)[1]. In this paper we generalize these results to all graphlets of density ≥ 0.85, up to a node size of 14 (for definitions see Section 2), as well as all complete bipartite graphlets of density ≥ 0.55 up to a node size of 10. The reason that we limited the results to graphlets of density ≥ 0.85 is that for lower densities the number of these occurrences seems to be extremely high and the algorithm took alot of time to give results. Based on these counts we present graphlet statistics for (1) two of the best studied PPI networks: Yeast (eukaryotic) and E.Coli (prokaryotic) as well as (2) for random networks that are obtained through GDM and the GRGM models, whose parameters are set to emulate the growth of these PPI networks. Our main observation is that for the two organisms considered, the motif statistics were in accordance with their complexity: the Yeast network, in comparison to the E.coli network, contained (relatively) many more denser graphlets than sparser graphlets. We also observe that dense motif statistics of the Yeast network was highly divergent from that of the GDM whose parameters were set to emulate its growth as closely as possible. Finally, we prove that low-dimensional GRGM cannot generate complete, bipartite graphs - which abundantly occur in the PPI networks we considered. These observations may help resolve the inconclusive discussion on the suitability of random network generators in emulating the evolution of PPI networks.

2. Inference of Densely Connected and Bipartite Subgraphs

An undirected graph $G = (V, E)$ is said to be a *supergraph* of G' if G' is an induced graphletof G. As usual, a graph is defined to be *bipartite* if $V = V_1 \dot{\cup} V_2$ such that $E \cap ((V_1 \times V_1) \cup (V_2 \times V_2)) = \emptyset$ and it is moreover called *complete* if $E = (V_1 \times V_2) \cup (V_2 \times V_1)$. A $K_{n,m}$ is a complete, bipartite graph such that $|V_1| = n, |V_2| = m$.

We define the *density* $d(G)$ of a graph G as the number of edges divided by the number of possible edges $d(G) = \frac{|E|}{\binom{|V|}{2}} = \frac{2|E|}{|V|(|V|-1)}$. G is said to be α-*dense* if $d(G) \geq \alpha$ and *dense* in general if $\alpha = 0.5$. As usual, G is said to be *connected* if there is a path between any pair of nodes in G. G is said to be *densely connected* if it is dense and connected.

Given a biomolecular network $G = (V, E)$ the driving question is to count all its induced densely connected and complete bipartite subgraphs. Note that inference of all densely connected subgraphs is $\mathcal{NP}$-hard which can be shown by a straightforward reduction from the max-clique problem which is $\mathcal{NP}$-complete[8]. Therefore, one has to screen all $2^{|V|}$ subgraphs in the worst case. Help comes from the following insights:

(i) For $\alpha \geq 0.5$, in each α-dense graph of node size n there is an induced α-dense graphletof size $n - 1$

(ii) Each induced graphlet of a complete bipartite graph is a complete bipartite graph.

While proving (ii) is straightforward, (i) is based on some more subtle arguments that have been presented elsewhere[4]. As a consequence of a combination of $(i), (ii)$ one can employ a search strategy starting with induced 2-node subgraphs and iteratively not considering supergraphs that contradict (i) and/or (ii).

Thanks to the peculiarities of PPI networks (sparseness, scale-freeness), by employing a search strategy that incorporates this core idea, the inference problem becomes tractable.

3. Models for PPI Network Emulation

3.1. *The Generalized Duplication Model*

The duplication model grows iteratively in discrete time steps. It starts with an arbitrary connected network $G(t_0)$, of node size t_0. In iteration $t > t_0$, one node, denoted as v_t, is added to $G(t-1)$, the network resulting from iteration $t-1$, as follows:

A node $w \in G(t-1)$ is picked uniformly at random and then "duplicated" by creating a new node v_t that is connected to all neighbors of w but not to w itself. Subsequently,

(1) Edges (u, v_t) (where u is a neighbor of w) are deleted with probability p,

(2) Edges (u, v_t) are created with probability $r/(t-1)$ where u runs through all nodes in the network. Resulting parallel edges are merged if necessary.

In this process, p and r are fixed parameters.

3.2. *Geometric Random Graphs*

A geometric random graph (GRG) is created by drawing points uniformly at random from some restricted area (e.g. the hypercube) as nodes and, given some fixed threshold r, connecting two nodes v_i, v_j by an edge if $||v_i - v_j|| < r$ where $||.||$ is the Euclidean norm. The dimension of the hypercube is referred to as the dimension of the graph. GRGs have been suggested as a model for emulating PPI networks[12,6]. Although geometric random graphs can successfully capture a couple of PPI network features we would like to outline that GRGs cannot generate induced, complete bipartite subgraphs in the following. However, such subgraphs are crucial ingredients of PPI networks[3].

Theorem 3.1. *A* 2- *resp.* 3-*dimensional GRG does not contain a* $K_{2,3}$ *resp.* $K_{3,3}$ *as an induced subgraph.*

As we will demonstrate (see the results Section 4) that PPI networks contain significant amounts of such induced bipartite subgraphs in the following, this will rule out low-dimensional GRGs as a model that is suited to capture important local topological features of PPI networks. Moreover, as inspired by the theorem's proof, we conjecture that there is no $K_{n,2}$ in $(n-1)$-dimensional and no $K_{n,3}$ in n-dimensional GRGs. We are currently close to finishing respective results for

$n = 4$ and, indeed, we couldn't observe $K_{4,2}$'s resp. $K_{4,3}$'s in the 3- resp. 4-dimensional GRGs we generated.

3.3. *Notations*

In the following, for $X, Y \in \mathbb{R}^d$ let

$$\overline{XY} := \{v \in \mathbb{R}^d : v = \lambda X + (1-\lambda)Y : \lambda \in [0,1]\}$$

be the line segment that connects X and Y and

$$\overrightarrow{XY} := \{v \in \mathbb{R}^d : v = X + \lambda(Y - X), \lambda \geq 0\}$$

be the half ray that leaves X in direction of Y. For $X \in \mathbb{R}^d$ let

$$U_r(X), S_r(X), B_r(X) := \{v \in \mathbb{R}^d : ||v - X||\ \{<, =, \leq\}\ r\}$$

be the open ball, the (hyper)sphere and the closed ball with radius r around X. Note that for $d = 2$, $S_r(X)$ is just a circle around X with radius r.

3.4. *Proof of theorem 3.1*

We prepare the proof with two essential lemmata.

Lemma 3.1. *Let $P, A, B, C \in \mathbb{R}^2$ such that*

$$||P - A||, ||P - B||, ||P - C|| < r \leq ||A - B||, ||A - C||, ||B - C|| \qquad (1)$$

Then, for $s \geq \max\{||P - A||, ||P - B||, ||P - C||\}$, it holds that

$$U_r(A) \cap U_r(B) \cap U_r(C) \subset U_s(P). \qquad (2)$$

Proof. Obviously, if A, B, C are located on one line, $U := U_r(A) \cap U_r(B) \cap U_r(C)$ is empty such that there is nothing to prove. Therefore, we can assume that the affine dimension of A, B, C is 2. In combination with some elementary geometric arguments, this implies that $\overrightarrow{PA}, \overrightarrow{PB}, \overrightarrow{PC}$ divide $\mathbb{R}^2$ into three sections such that the section S_{XY} which is bounded by $\overrightarrow{PX}, \overrightarrow{PY}$ does not contain Z where $X, Y, Z \in \{A, B, C\}$ (X, Y, Z pairwise different). Hence

$$\mathbb{R}^2 = S_{AB} \cup S_{AC} \cup S_{BC} \quad \text{and} \quad A \notin S_{BC}, B \notin S_{AC}, C \notin S_{AB}. \qquad (3)$$

We will show that

$$U_r(A) \cap S_{BC},\ U_r(B) \cap S_{AC},\ U_r(C) \cap S_{AB}\ \subset\ U_s(P) \qquad (4)$$

which yields

$$U \overset{(3)}{=} U \cap (S_{AB} \cup S_{AC} \cup S_{BC}) = (U \cap S_{AB}) \cup (U \cap S_{AC}) \cup (U \cap S_{BC})$$
$$\subset (U_r(C) \cap S_{AB}) \cup (U_r(B) \cap S_{AC}) \cup (U_r(A) \cap S_{BC}) \overset{(4)}{\subset} U_s(P)$$

and therefore establishes (2).

In order to show (4), we can restrict our attention to showing $U_r(C) \cap S_{AB} \subset U_s(P)$ as the remaining two cases follow from arguments that are completely analogous. If we can show that

$$S_r(C) \cap S_{AB} \subset U_s(P). \tag{5}$$

we will be done with the proof by the following argumentation. Let $D \in U_r(C) \cap S_{AB}$. Consider $\overrightarrow{PD}$. By definition of S_{AB} as a section we have

$$\overrightarrow{PD} \subset S_{AB}. \tag{6}$$

Note that every ray starting within $U_r(C)$ will finally hit the boundary of $U_r(C)$ which is $S_r(C)$. Therefore, while traveling from $D \in U_r(C)$ on $\overrightarrow{PD}$ away from P, one will hit $S_r(C)$. Denote by E the point of that intersection. We compute

$$E \in S_r(C) \cap \overrightarrow{PD} \overset{(6)}{\subset} S_r(C) \cap S_{AB} \overset{(5)}{\subset} U_s(P).$$

However, by construction, D is at least as close to P as E, so $D \in U_s(P)$ with which we have completed the proof.

In order to finally show (5) observe that, because of (1), $P \in U_r(C)$, that is, P is inside of $S_r(C)$, whereas A, B are not inside. Therefore, $S_r(C)$ intersects $\overrightarrow{PA}$ resp. $\overrightarrow{PB}$ between P and, at the latest (maybe just there), A resp. B. Therefore, by definition of s,

$$S_r(C) \cap \overrightarrow{PA},\ S_r(C) \cap \overrightarrow{PB} \ \subset\ B_s(P). \tag{7}$$

However, for $F \in S_r(C) \cap \overrightarrow{PC}$ we have that, as all $P, C, F \in \overrightarrow{PC}$ (*),

$$||F - P|| \overset{(*)}{=} ||P - C|| + ||C - F|| = ||P - C|| + r > r \geq s$$

where $||P - C|| > r$ follows from (i) which implies $P \neq C$ whereas the last inequation follows from the definition of s. This translates to

$$S_r(C) \cap \overrightarrow{PC} \ \not\subset\ B_s(C) \tag{8}$$

Note that two non-concentric circles (here: $S_r(C), S_s(P)$) can intersect at most two points. In our case, these two points establish the transitions of $S_r(C)$ from being inside $B_s(P)$ to being outside $B_s(P)$. Combining this insight with (7,8) implies that, within S_{AB}, $S_r(C)$ is contained in $B_s(P)$ which establishes (5). □

In the following we will write R_x, R_y, R_z for the coordinates of $R = (R_x, R_y, R_z) \in \mathbb{R}^3$.

Lemma 3.2. *Let $P, Q, A, B, C \in \mathbb{R}^3$ such that*

$$A_z = B_z = C_z = 0 \quad \textit{and} \quad P_z \geq Q_z \geq 0 \tag{9}$$

$$||P-A||, ||P-B||, ||P-C||, ||Q-A||, ||Q-B||, ||Q-C|| < r \tag{10}$$

$$||A-B||, ||A-C||, ||B-C|| \geq r. \tag{11}$$

Then, for $s \geq \max\{||P-A||, ||P-B||, ||P-C||\}$, it holds that

$$Q \in U_s(P). \tag{12}$$

Proof. Intuitively speaking, this lemma is trying to show that if three points in two-dimensional space are in proximity of two other points, then these two points have to be proximate to each other.

Let $P' = (P_x, P_y, 0), Q' = (Q_x, Q_y, 0)$ be the projection of P and Q onto the $x-y$-space. Combining (9) and (10) implies that also

$$||P'-A||, ||P'-B||, ||P'-C||, ||Q'-A||, ||Q'-B||, ||Q'-C|| < r. \tag{13}$$

Hence P', A, B, C satisfy the conditions of lemma 3.1 and applying the lemma yields

$$||P'-Q'|| \leq \max\{||P'-A||, ||P'-B||, ||P'-C||\}. \tag{14}$$

Wlog. let A such that $||P'-Q'|| \leq ||P'-A||$. We compute

$$\begin{aligned} ||P-Q|| &= \sqrt{(P_x-Q_x)^2 + (P_y-Q_y)^2 + (P_z-Q_z)^2} \\ &= \sqrt{||P'-Q'||^2 + (P_z-Q_z)^2} \overset{(14)}{\leq} \sqrt{||P'-A||^2 + (P_z-Q_z)^2} \\ &\overset{(ii)}{\leq} \sqrt{||P'-A||^2 + P_z^2} = ||P-A||. \end{aligned}$$

□

We are now in position to prove theorem 3.1.

Proof. [Th. 3.1] We assume the contrary in both cases.

For the first part let $A, B, C \in \mathbb{R}^2$ and $P, Q \in \mathbb{R}^2$ be the partitions of the sampled $K_{3,2}$. By definition of a $K_{3,2}$ in a GRG, P, A, B, C satisfy (1) in lemma 3.1 where r is the threshold of the 2-dimensional GRG. Applying lemma 3.1 yields $||P-Q|| < s \leq r$ which is a contradiction to having an edge between P and Q.

For the second part let $P, Q, R \in \mathbb{R}^3$ and $A, B, C \in \mathbb{R}^3$ be the two partitions of the sampled $K_{3,3}$. Observe that any three points in $\mathbb{R}^3$ lie on a plane. By necessarily rotating all points around the origin (which is an orthogonal transformation hence norm preserving) we can assume wlog. that A, B, C lie in the $x-y$-space of $\mathbb{R}^3$. By necessarily reflecting P, Q, R at the $x-y$-space which, again, is a norm preserving transformation, we can moreover assume that, for two of the P, Q, R, the z-coordinates are non-negative. Wlog. $P_z \geq Q_z \geq 0$. By definition of a $K_{3,3}$ in a 3-dimensional GRG, this establishes (9,10,11) of lemma 3.2 where r is the threshold of the 3-dimensional GRG. Applying lemma 3.2 to P, Q, A, B, C yields $Q \in U_s(P)$ which yields $||P-Q|| < s \leq r$. This is a contradiction to that there is no edge between P and Q. □

4. Results

We present our results on dense graphlet statistics for both Yeast and E.coli PPI networks for graphlet densities in the range $[0.85, 1.00]$ and number of nodes in the range $[7, 14]$.[a] We also present similar dense motif statistics for random graphs generated by both 3-dimensional GRGM (unit cube) and GDM which are set to generate identical number of nodes and edges to the Yeast PPI network[13] [b], i.e. $|V| \approx 4900$ and $|E| \approx 17000$.[c] Due to limitations on computational resources, we were not able to count graphlets of larger sizes or smaller density. In our experiments we have used the network available from DIP [13] for Yeast and E.coli. Yeast PPI network has around 4900 nodes and 17000 edges, and E.coli network has around 1441 nodes and 5871 edges. By considering the distribution (fraction) of dense graphlets (between range $[0.85, 1.00]$) for each species, we are trying to eliminate the effect caused by difference in average degree between these two species.

In Figure 1 we depict for each $n \in \{7, \ldots, 12\}$ (the number of nodes in the graphlet), how the proportion of all graphlets with k edges among all dense graphlets (with density ≥ 0.85) vary with respect to k - for all four PPI networks considered. Note that we give only proportional distributions here as the Yeast and the E.coli networks have different number of edges and nodes and thus it is not meaningful to compare absolute figures.

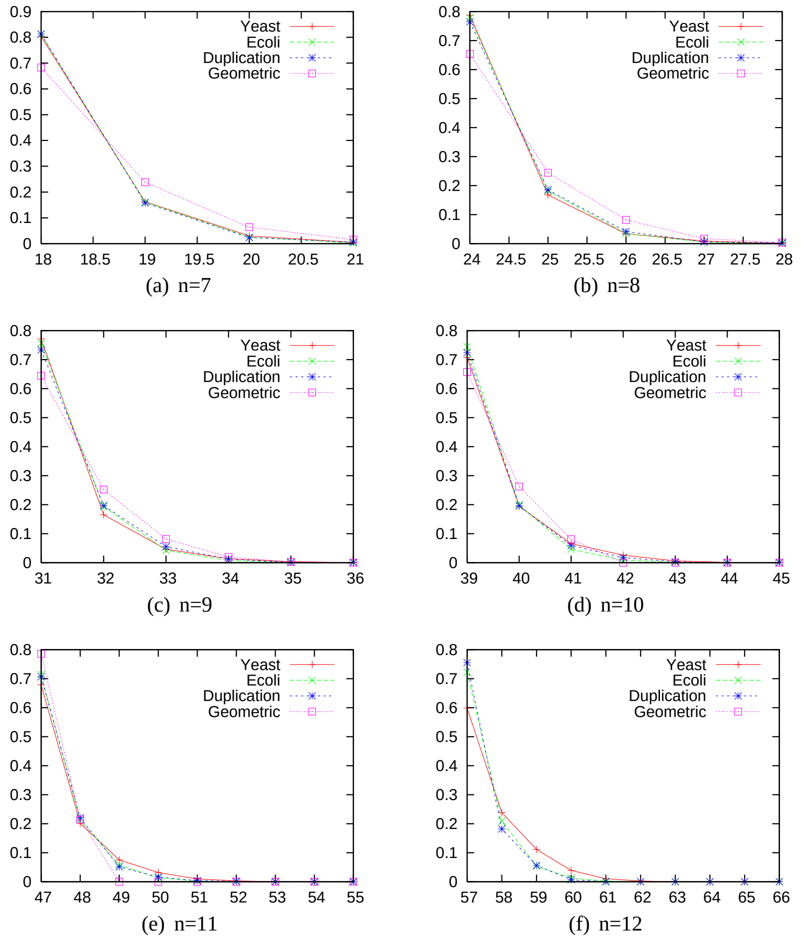
It is possible to make a number of observations on Figure 1:

(1a) The fractional distribution of dense graphlets in GRGM is significantly different from that of the other networks: for example, it contains no graphlets with density ≥ 0.85 for $n = 12, 13$ and 14 and no graphlets with density ≥ 0.89 (i.e.

[a] We remind the reader that the density of a graphlet with n nodes is the ratio of the number of edges in the graphlet and $n(n-1)/2$, the maximum number of edges possible between the nodes of the graphlet; a density of 1 indicates a clique.

[b] The DIP release date for Yeast and E.coli PPI networks is July 7, 2007.

[c] This is possible by setting the radius of the GRGM to an appropriate value and setting $p = 0.365$ and $r = 0.12$ in the GDM as per[7,1].



(a) n=7

(b) n=8

(c) n=9

(d) n=10

(e) n=11

(f) n=12

Figure 1. Here we demonstrate for each $n \in \{7, 8, 9, 10, 11, 12\}$ the fraction of graphlets in each PPI network which have k edges, among all graphlets with density ≥ 0.85. The specific PPI networks are depicted with the following colors: Yeast (Red), E.coli (Green), GDM (Blue) and GRGM (Purple). For example, plot (a) indicates that, the graphlets in the E.coli PPI network (Green) with 7 nodes and 18 edges, cumulatively, form 80% of all of its dense graphlets with density ≥ 0.85.

with 49 edges or more) for $n = 11$. This, together with the proof that we provide in Section 3.4, this observation seems to support a negative answer to the question (i-a) in Section 1.1.

(1b) The fractional distribution of GDM largely agrees with the two PPI networks for $n \leq 12$. However, GDM fails to generate any dense network (with density ≥ 0.85) with $n = 13$ or 14.

(2) The fractional distribution (of dense graphlets) in Yeast and E.coli are quite similar for smaller values of n. However for $n = 11$ and especially for $n = 12$, the fraction of graphlets in Yeast increase with density. For example, for $n = 12$, the fraction of graphlets with 59 edges is ~ 0.1 in Yeast whereas it is ~ 0.05 in E.coli. This observation seems to provide a positive answer to question (ii) in Section 1.1 (although we believe much more study should be done regarding this question).

In Figure 2, we depict how the total number of dense graphlets (with density ≥ 0.85) with n nodes ($n = \{3, \ldots, 14\}$) vary as a function of n - for the Yeast PPI network as well as the specific networks generated by GRGM and GDM whose parameters were set to emulate the Yeast network.

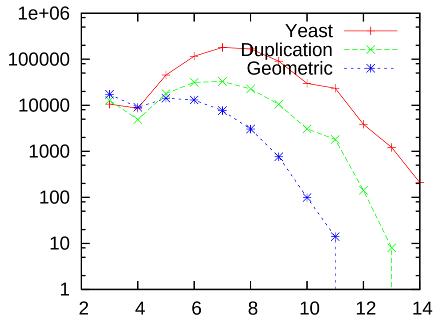


Figure 2. Total number of dense subgraphs (with density ≥ 0.85) with n nodes - in the Yeast PPI network as well as networks generated by GRGM and GDM - whose parameters are set to emulate the Yeast PPI network. The specific colors used are GRGM (Blue), GDM (Green) and Yeast (Red).

As can be seen in Figure 2, there is a wide gap between the total numbers of dense graphlets in the the Yeast PPI network and the random networks generated by GDM and GRGM. Although the number of dense graphlets for $n = 6$ is consistent with an earlier study [7], the figure 2 shows substantially difference for $n > 6$ between GDM (or GRGM) and Yeast PPI network, especially for $n \geq 8$, where there is a 7-fold (or 50-fold) difference respectively. More drastically GRGM includes no dense graphlets with $n = 12$ nodes and GDM includes no dense graphlets with $n = 14$ nodes. Once again, Figure 2 seems to support a negative answer to question (i-a) in Section 1.1: i.e. GRGM is not suitable for emulating the growth of PPI networks. The figure also seems to imply (somewhat less strongly) a negative answer to question (i-b), i.e., the GDM used in this paper does not capture the dense graphlet distribution of the Yeast PPI network. We leave the possibility of how (or if) we can modify the seed network or the GDM itself so as to better capture the distribution of denser graphlets of size bigger than 6 (see [7]) via a duplication oriented model to future study. Our final results are on the fractional distributions of all $K_{n,n}$'s and all $K_{n,n-1}$'s (which are all 0.55-dense complete bipartite graphs) up to $n = 5$ in each of the PPI networks we considered.

Bipartite Graph	Ecoli	Yeast	Duplication	Geometric
$K_{2,3}$	2685054	498844	337218	153
$K_{3,3}$	2188868	376186	23311	0
$K_{3,4}$	11103153	1677626	21623	0
$K_{4,4}$	5155489	852301	519	0
$K_{4,5}$	13561155	2077675	129	0
$K_{5,5}$	1125496+	659614	2	0

In the above table, it can be seen that in the E.Coli and the Yeast PPI networks, complete bipartite graphlets are abundant. However, as can be deduced from theorem 3.1, the GRGM cannot generate $K_{n,m}$'s for $n, m \geq 3$. Our experiments confirm this finding: there are no $K_{n,n}$'s and no $K_{n,n+1}$'s in the GRGM network for $n \geq 3$.

References

1. N. Alon, P. Dao, I. Hajirasouliha, F. Hormozdiari and S.C. Sahinalp, *Proceedings of the ISMB 2008*, (2008).
2. U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits* (2006).
3. K. Baskerville and M. Paczuski, *Physical Review E* **74**, 051903 (2006).
4. R. Colak, F. Moser, A. Rafiey and M. Ester, *Tech Report*, Simon Fraser University, (2008).
5. J. Grochow and M. Kellis, *Proc. RECOMB 2008*, 92, (2008).
6. D.J. Higham, M. Rasajski and N. Przulj, *Bioinformatics* **24**, 1093 (2008).
7. F. Hormozdiari, P. Berenbrink, N. Przulj and S.C. Sahinalp, *PLoS Computational Biology* **3**, e118 (2007).
8. R.M. Karp, in *Complexity of Computer Computations, Plenum Press*, 85, 1972. *Theoretical Computer Science* **369**, 234 (2006).
9. M. Middendrof, E. Ziv, and C. Wiggins, *PNAS* **102**, 9 (2005).
10. R. Milo, S. Shen-Orr, S. Itzkovski, N. Kashtan, D. Chklovskii and U. Alon, *Science* **298**, 824 (2002).
11. ME. Newman, SH. Strogatz, DJ. Watts, *Phys Rev E Stat Nonlin Soft Matter Phys.* (2001).
12. N. Przulj, D.G Corneil and I. Jurisica, *Bioinformatics* **20**, 3508 (2004).
13. L. Salwinski, C.S. Miller, A.J. Simith, F.K. Pettit, J.U. Bowie and D. Eisenberg, *Nucleic Acid Research* **32**, Database issue:D449-52 (2004).