Counting five-node subgraphs

Steve Lawford

Counting five-node subgraphs

Steve Lawford
Data, Economics and Interactive Visualization (DEVI) group, ENAC (University of Toulouse),
7 avenue Edouard Belin, CS 54005, 31055, Toulouse, Cedex 4, France
Email: steve.lawford@enac.fr

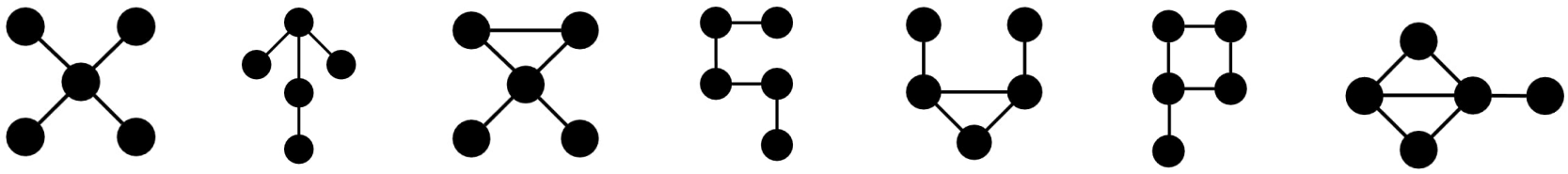
Abstract

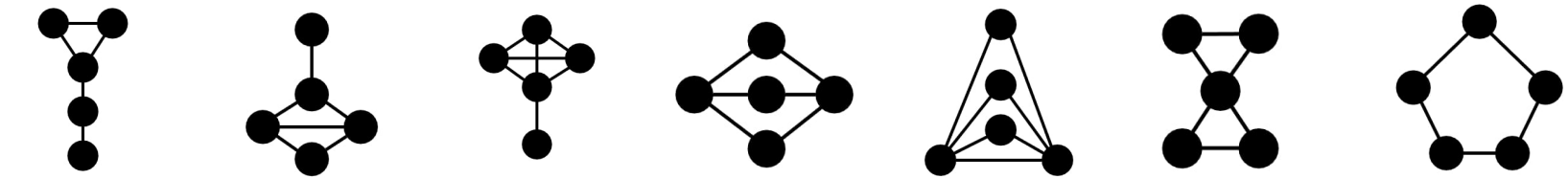
We propose exact count formulae for the 21 topologically distinct non-induced connected subgraphs on five nodes, in simple, unweighted and undirected graphs. We prove the main result using short and purely combinatorial arguments that can be adapted to derive count formulae for larger subgraphs. To illustrate, we give analytic results for some regular graphs, and present a short empirical application on real-world network data. We also discuss the well-known result that induced subgraph counts follow as linear combinations of non-induced counts.

1 Introduction

Networks are a fundamental tool for modelling the topological structure of complex systems. They are of considerable theoretical interest, and have practical applications in a wide variety of fields, including biology, statistical physics, and social science [3, 9, 13, 20, 28, 32]. An important problem is the enumeration of small connected induced or non-induced subgraphs (graphlets) on a network, a variant of the classical subgraph isomorphism problem in theoretical computer science. For example, subgraph counts are used for network classification, and to determine the statistical significance of small topological structures that arise frequently in real-world networks, where they may have specific functional roles [6, 7, 12, 16, 19, 22, 23, 25, 26, 34]. Analytical exact count formulae have been a subject of theoretical research since at least the early 1970s [15], although it can be computationally expensive to apply them, and this becomes exponentially hard as the number of nodes in the graphlet increases (combinatorial explosion). For this reason, there has been a great deal of work on the design of efficient exact and approximate sampling algorithms, with applications to increasingly massive datasets (the first subgraph counting algorithm appeared in [18]; for recent work, see references in [8] and the comprehensive survey by [31]). However, there are almost no complete analytical treatments of exact subgraph counting on five nodes, despite the theoretical insights that can be gained from such formulae, and their usefulness in applications such as [2, 4, 24].

We fill this gap here by providing exact count formulae for the 21 topologically distinct non-induced connected subgraphs on five nodes (Table 1), in simple, unweighted and undirected graphs. We use purely combinatorial and elementary techniques which give rise to correspondingly intuitive count formulae that are very convenient for theoretical work. In a number of cases, we find quite different formulations of known results. The techniques that are used to derive exact results on the eight connected subgraphs with three or four nodes are very well known [2, 13, 14]. On the other hand, the number of connected graphs increases very rapidly in the number of nodes, and a complete treatment even of the 112 subgraphs on six nodes is currently out of reach, although exact count formulae can certainly be found for individual graphlets with six or more nodes (see A001349 of the Online Encyclopedia of Integer Sequences (`http://oeis.org/A001349`) for the number of graphlets on n nodes).





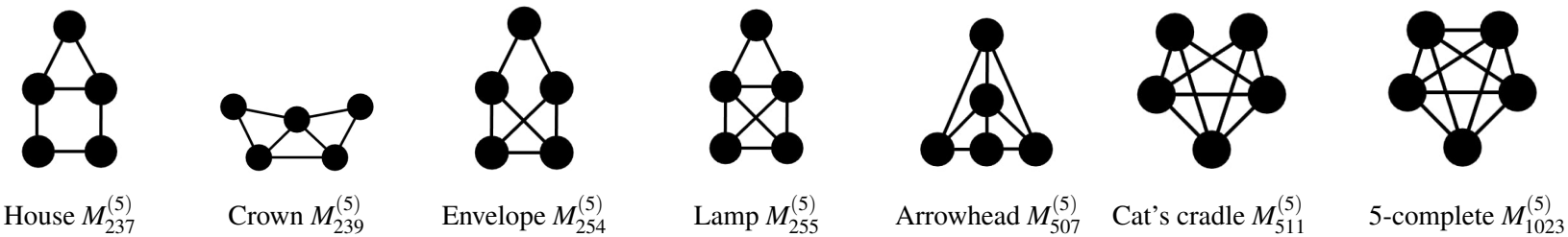


Table 1: *The 21 topologically distinct connected subgraphs on five nodes, denoted $M_a^{(b)}$ (see Section 1.1 for notation).*

The paper is organized as follows. In Section 2 we present the main result and discuss the key strategies that are used in the proofs. In Section 3 we specialize selected count formulae to three classical regular graphs, namely the complete graph K_n, the complete N-partite graph $K_{n_1,n_2,\ldots,n_N}$, and the regular ring lattice, and use these results to develop combinatorial intuition in each case. We then show how alternative methods of proof can lead to very different formulae and, sometimes, to confusion or error. We also illustrate how all of our count formulae can be implemented in a practical manner on a small real-world dataset. Section 4 concludes. In Appendix A we discuss the well-known result that induced subgraph counts follow immediately as linear combinations of non-induced counts.

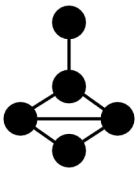
1.1 Notation and preliminaries

Let $G=(V,E)$ be a graph with node set V and edge set E. We write $n=|V|$ and $m=|E|$ for the number of nodes and edges of G. Let g be the $n\times n$ adjacency matrix corresponding to G, with representative element $(g)_{ij}$, that takes value one when an edge is present between nodes i and j, and zero otherwise. We use $(i,j)\in E$ to denote an edge between nodes i and j, and say that they are directly-connected. A graph is simple and unweighted if $(g)_{ii}=0$ and $(g)_{ij}\in\{0,1\}$, and undirected if $(g)_{ij}=(g)_{ji}$. In some proofs, we use $(i,j)^\star$ to denote "(i,j) and (j,i)", when the direction of the edge is important. A walk between nodes i and j is a sequence of edges $\{(i_r,i_{r+1})\}_{r=1,\ldots,R}$ for which $i_1=i$ and $i_{R+1}=j$, and a path is a walk over distinct nodes. A graph is connected if there is at least one path between any pair of nodes i and j. We use $\Gamma_G(i)=\{j:(i,j)\in E\}$ to denote the neighbourhood of node i in G, with cardinality equal to the node degree $k_i=\sum_j(g)_{ij}$, and let $P(k)$ be the distribution of node degrees. We denote the common neighbourhood of nodes i and j by $S(i,j)=\Gamma_G(i)\cap\Gamma_G(j)$, with cardinality $\#(S)=(g^2)_{ij}$. Let $G'=(V',E')\subseteq G$ denote a subgraph of G, such that $V'\subseteq V$ and $E'\subseteq E$. If $G'\subseteq G$ and all the edges $(i,j)\in E$ such that $i,j\in V'$ are in E', then G' is an induced subgraph of G (otherwise it is "non-induced"). Special graphs are the complete graph on n nodes, K_n, that has all possible edges, and the Erdős-Rényi random graph $G(n,p)$ with node set $V=\{1,\ldots,n\}$ and edges that arise independently with constant probability p. We use the notation $M_a^{(b)}$ of [2, 24] to refer to a specific non-induced graphlet, where b is the number of nodes in the subgraph and a is the decimal representation of the smallest binary number derived from a row-by-row reading of the upper triangles of each adjacency matrix g across the set of all isomorphic graphlets on the same b nodes:

$$a=\sum_{i=1}^{b-1}\sum_{j=i+1}^{b}2^{\binom{b-i}{2}+(b-j)}(g)_{ij}.$$

Let $\widetilde{M}_a^{(b)}$ be an induced subgraph. The non-induced and induced subgraph counts in G are denoted $|M_a^{(b)}|$ and $|\widetilde{M}_a^{(b)}|$.

Example 1.1. The notation itself fully defines the topology of each graphlet, and enables us to state clear and unambiguous count formulae later on. Consider the spinning top graphlet represented by:



The set of decimal representations of the adjacency matrix is given by

$$\begin{aligned}\mathscr{A}=\{&119,125,126,175,187,190,231,249,287,315,317,343,378,399,444,461,\\&462,467,470,473,483,485,490,500,543,559,567,605,622,667,694,711,717,\\&723,730,732,745,746,753,756,811,821,839,846,857,858,867,873,876,882,\\&884,903,918,921,924,933,938,940,945,946\}.\end{aligned}$$

The smallest and largest elements of $\mathscr{A}$ have binary representations $119_{10} = 0001110111_2$ and $946_{10} = 1110110010_2$ respectively. We denote the spinning top by $M_{119}^{(5)}$, where the nodes are labelled as follows: i_1 (the degree 1 node), i_2 (the degree 2 node), i_3 and i_4 (the degree 3 nodes that are not directly-connected to i_1, in any order), and i_5 (the degree 3 node that is directly-connected to i_1).

2 The main result

Our main result presents analytic count formulae for all non-induced five node subgraphs, in terms of functions of the node degrees for various subsets of nodes in the graph, powers of the adjacency matrix, powers of the adjacency matrix formed by removal of a given node, and smaller non-induced subgraphs. A full combinatorial proof follows the statement of Theorem 2.1.

Theorem 2.1 (Count formulae for non-induced graphlets on five nodes)**.**

$$|M_{75}^{(5)}| = \sum_{i:k_i>3} \binom{k_i}{4} = \frac{1}{24} \sum_{i:k_i>3} k_i(k_i-1)(k_i-2)(k_i-3). \tag{1}$$

$$|M_{77}^{(5)}| = \sum_{\substack{(i,j)^\star \in E \\ k_i>2}} \binom{k_i-1}{2}(k_j-1) - 2|M_{15}^{(4)}| = \frac{1}{2} \sum_{\substack{(i,j)^\star \in E \\ k_i>2}} (k_i-1)(k_i-2)(k_j-1) - 2|M_{15}^{(4)}|. \tag{2}$$

$$|M_{79}^{(5)}| = \frac{1}{2} \sum_{i:k_i>3} (g^3)_{ii} \binom{k_i-2}{2} = \frac{1}{4} \sum_{i:k_i>3} (g^3)_{ii}(k_i-2)(k_i-3). \tag{3}$$

$$|M_{86}^{(5)}| = \frac{1}{2} \sum_{i,j:i\neq j} (g^4)_{ij} - 2\,|M_3^{(3)}| - 9\,|M_7^{(3)}| - 3\,|M_{11}^{(4)}| - 2\,|M_{13}^{(4)}| - 2\,|M_{15}^{(4)}|. \tag{4}$$

$$|M_{87}^{(5)}| = \sum_{\substack{(i,j)\in E \\ k_i>2,\, k_j>2}} (g^2)_{ij}(k_i-2)(k_j-2) - 2\,|M_{31}^{(4)}|. \tag{5}$$

$$|M_{94}^{(5)}| = \sum_{\substack{i,j:i\neq j \\ k_i>2}} \binom{(g^2)_{ij}}{2}(k_i-2) - 2\,|M_{31}^{(4)}| = \frac{1}{2} \sum_{\substack{i,j:i\neq j \\ k_i>2}} (g^2)_{ij}((g^2)_{ij}-1)(k_i-2) - 2\,|M_{31}^{(4)}|. \tag{6}$$

$$|M_{95}^{(5)}| = \sum_{\substack{i,j:i\neq j \\ k_i>3}} \binom{(g^2)_{ij}(g)_{ij}}{2}(k_i-3) = \frac{1}{2} \sum_{\substack{(i,j)^\star \in E \\ k_i>3}} (g^2)_{ij}((g^2)_{ij}-1)(k_i-3). \tag{7}$$

$$|M_{117}^{(5)}| = \frac{1}{2} \sum_{(i,j)^\star \in E} (g^3)_{ii}(k_j-1) - 6\,|M_7^{(3)}| - 2\,|M_{15}^{(4)}| - 4\,|M_{31}^{(4)}|. \tag{8}$$

$$|M_{119}^{(5)}| = \sum_{\substack{(i,j)\in E \\ k_i>2,\, k_j>2}} ((g^2)_{ij}-1) \sum_{\substack{r\in S(i,j) \\ k_r>1}} (k_r-2) - 12\,|M_{63}^{(4)}|. \tag{9}$$

$$|M_{127}^{(5)}| = \sum_{i:k_i>3} |M_{63}^{(4)}(g_{-i})|\,(k_i-3) = \frac{1}{6} \sum_{i:k_i>3} \mathrm{tr}(g_{-i}^3)(k_i-3). \tag{10}$$

$$|M_{222}^{(5)}| = \frac{1}{2} \sum_{\substack{i,j:i\neq j \\ k_i>2, k_j>2}} \binom{(g^2)_{ij}}{3} = \frac{1}{12} \sum_{\substack{i,j:i\neq j \\ k_i>2, k_j>2}} (g^2)_{ij}((g^2)_{ij}-1)((g^2)_{ij}-2). \tag{11}$$

$$|M_{223}^{(5)}| = \frac{1}{2} \sum_{\substack{i,j:i\neq j \\ k_i>3, k_j>3}} \binom{(g^2)_{ij}(g)_{ij}}{3} = \sum_{\substack{(i,j)\in E \\ k_i>3, k_j>3}} (g^2)_{ij}((g^2)_{ij}-1)((g^2)_{ij}-2). \tag{12}$$

$$|M_{235}^{(5)}| = \sum_{i:k_i>2} \binom{\frac{1}{2}(g^3)_{ii}}{2} - 2\,|M_{31}^{(4)}| = \frac{1}{8} \sum_{i:k_i>3} (g^3)_{ii}((g^3)_{ii}-2) - 2\,|M_{31}^{(4)}|. \tag{13}$$

$$|M_{236}^{(5)}| = \frac{1}{10}\left(\mathrm{tr}(g^5) - 30\,|M_7^{(3)}| - 10\,|M_{15}^{(4)}|\right). \tag{14}$$

$$|M_{237}^{(5)}| = \sum_{(i,j)\in E} (g^3)_{ij}(g^2)_{ij} - 9\,|M_7^{(3)}| - 2\,|M_{15}^{(4)}| - 4\,|M_{31}^{(4)}|. \tag{15}$$

$$|M_{239}^{(5)}| = \sum_{i:k_i>3} |M_{13}^{(4)}(g_{-i})| = \sum_{i:k_i>3} \left(\sum_{(j,r)\in E(\Gamma_G(i))} (k_j-1)(k_r-1) - 3\,|M_7^{(3)}(g_{-i})| \right). \tag{16}$$

$$|M_{254}^{(5)}| = \sum_{\substack{(i,j)\in E \\ k_i>2, k_j>2}} \sum_{\substack{r,q\in S(i,j) \\ r\neq q, k_r>2, k_q>2}} ((g^2)_{rq}-2). \tag{17}$$

$$|M_{255}^{(5)}| = \frac{1}{2} \sum_{i:k_i>3} |M_{15}^{(4)}(g_{-i})| = \frac{1}{4} \sum_{i:k_i>3} \sum_{r:k_r>2} (g_{-i}^3)_{rr}(k_r-2). \tag{18}$$

$$|M_{507}^{(5)}| = \sum_{i:k_i>3} |M_{30}^{(4)}(g_{-i})| = \frac{1}{8} \sum_{i:k_i>3} \left(\mathrm{tr}(g_{-i}^4) - 2\,m(g_{-i}) - 4\,|M_3^{(3)}(g_{-i})|\right). \tag{19}$$

$$|M_{511}^{(5)}| = \frac{1}{3} \sum_{i:k_i>3} |M_{31}^{(4)}(g_{-i})| = \frac{1}{6} \sum_{i:k_i>3} \left(\sum_{\substack{r,s:r\neq s \\ r>2, s>2}} \binom{(g_{-i}^2)_{rs}(g_{-i})_{rs}}{2} \right). \tag{20}$$

$$|M_{1023}^{(5)}| = \frac{1}{5} \sum_{i:k_i>3} |M_{63}^{(4)}(g_{-i})| = \frac{1}{120} \sum_{i:k_i>3} \sum_{\substack{j\in\Gamma_G(i) \\ k_j>2}} \mathrm{tr}(((g_{-i})_{-j})^3). \tag{21}$$

Proof of Theorem 2.1. We treat each subgraph separately. See Table 2 for the graphlets on three and four nodes.

(a) 5-star $|M_{75}^{(5)}|$: See [24, Proposition A.1, eqn. 18].

(b) 5-arrow $|M_{77}^{(5)}|$: See [24, Proposition A.1, eqn. 19].

(c) cricket $|M_{79}^{(5)}|$: The method of proof follows that used for the count of the tadpole $|M_{15}^{(4)}|$ in [2]. The cricket subgraph can be thought of as a triangle on nodes i, j and x, with the addition of two extra edges (i,y) and (i,z), where $k_i > 3$. The element $(1/2)(g^3)_{ii}$ is the number of triangles attached to node i, where the division by two corrects for double-counting due to the two possible directions of travel around the triangle. Hence, there are $(1/2)(g^3)_{ii}\binom{k_i-2}{2}$ crickets "centered on" node i. Result (3) follows immediately. See also [5, $n_G(H_7)$].

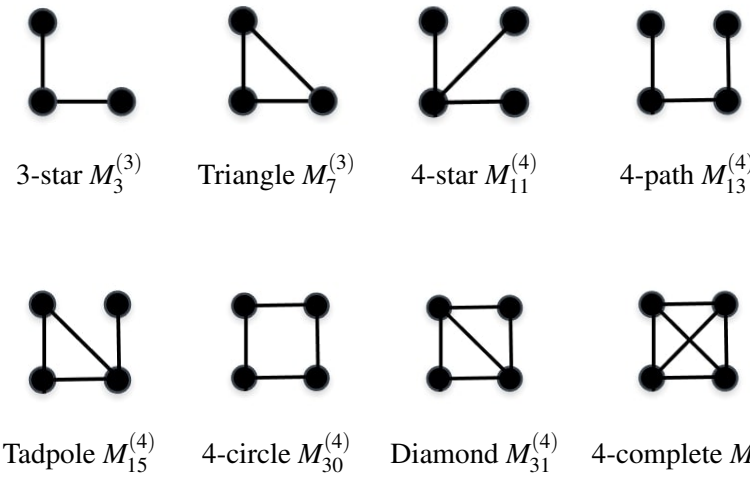


Table 2: *The 8 topologically distinct connected subgraphs on three and four nodes.*

(d) 5-path $|M_{86}^{(5)}|$: See [27, Theorem 4.1] and [24, Proposition A.1, eqn. 20].

(e) bull $|M_{87}^{(5)}|$: The proof uses similar ideas to the count of the 4-path $|M_{13}^{(4)}|$ in [2]. Consider any edge $(i,j) \in E$, as the central edge in a bull $\{(x,i),(x,j),(i,j),(i,y),(j,z)\}$. Given an edge (i,j), the term $(g^2)_{ij}$ is the number of triangles that contain the edge (i,j). Node i has $k_i - 2$ possible neighbours (for node y), and node j has $k_j - 2$ possible neighbours (for node z), where $k_i > 2$ and $k_j > 2$. There are $(k_i-2)(k_j-2)$ ways in which a neighbour of i can be paired with a neighbour of j, which gives a total of $\sum_{(i,j)\in E}(g^2)_{ij}(k_i-2)(k_j-2)$ candidate bulls across all possible central edges. This sum includes the unwanted case $y = z$, which forms a diamond with central edge (i,j). Since there are two paths of length two from node i to node j in a diamond, we subtract $2\,|M_{31}^{(4)}|$ to give result (5). See also [5, $n_G(H_8)$].

(f) banner $|M_{94}^{(5)}|$: Let i and j be the degree 3 and degree 2 nodes of a banner, such that $g_{ij} = 0$. Then, $(g^2)_{ij}$ gives the number of walks of length 2 between i and j. There are $\binom{(g^2)_{ij}}{2}$ pairs of these walks, and a banner is formed by linking node i to one of its (k_i-2) neighbours, denoted x. Hence, there are $\sum_{i,j:i\neq j}\binom{(g^2)_{ij}}{2}(k_i-2)$ candidate banners. This includes the unwanted case $x = j$ which gives a diamond. Since the central edge of a diamond has two endpoints, we subtract $2\,|M_{31}^{(4)}|$ to give result (6). See also [5, $n_G(H_9)$].

(g) stingray $|M_{95}^{(5)}|$: Let i and j be the degree 4 and degree 3 nodes in a stingray. Consider two triangles with a common edge (i,j). Given this common edge, $(g^2)_{ij}(g)_{ij}$ represents the number of walks of length 2 between i and j. A stingray is formed by any two of these triangles, with an additional edge attached to node i. This gives $\binom{(g^2)_{ij}(g)_{ij}}{2}(k_i-3)$ stingrays centered on node i. Summing across all pairs of nodes i and j gives the result (7).

(h) lollipop $|M_{117}^{(5)}|$: The proof uses similar ideas to that of the count of the nested 5-arrow $|M_{77}^{(5)}|$ in [24]. Consider any edge $(i,j) \in E$, as the central edge in a lollipop. Let i and j have degrees three and two respectively, and let node i be directly-connected to nodes x and z which form a triangle with i, and let node j also be directly-connected to node y. The term $(1/2)(g^3)_{ii}$ is the number of triangles that contain node i, where the division by two corrects for double-counting due to the two possible directions of travel around the triangle. Node j has $k_j - 1$ possible neighbours (for node y). Hence, there are $(1/2)(g^3)_{ii}(k_j-1)$ ways in which a triangle attached

to node i can be paired with a neighbour of node j, which gives a total of $(1/2)\sum_{(i,j)^\star\in E}(g^3)_{ii}(k_j-1)$ across all possible central edges, in both directions. This sum includes the unwanted cases $x=y$ and $z=y$, both of which form a diamond. Since four of the five edges of the diamond can be a candidate central edge (i,j) of a lollipop, we subtract $4\,|M_{31}^{(4)}|$. Furthermore, each edge of a triangle $M_7^{(3)}$, in both directions, and two edges of a tadpole $M_{15}^{(4)}$, in one direction, can be a candidate central edge (i,j) of a lollipop, and so we also subtract $6\,|M_7^{(3)}|$ and $2\,|M_{15}^{(4)}|$ to give result (8). See also [5, $n_G(H_{10})$].

(i) spinning top $|M_{119}^{(5)}|$: Given an edge $(i,j)\in E$, let $S(i,j)$ be the common neighbourhood of nodes i and j, with cardinality $\#(S)=(g^2)_{ij}$. Let r and r' be two members of S, each of which forms a triangle with a common edge (i,j). A spinning top is formed by linking the diamond on nodes i,j,r,r' with any other node x that is directly-connected to r (or r'). Given r, there are $(\#(S)-1)\sum_{r\in S(i,j),k_r\geq 2}(k_r-2)$ such spinning tops. We subtract $12\,|M_{63}^{(4)}|$ to correct for twice each edge in the unwanted case $x=r'$ (or $x=r$), and result (9) follows.

(j) kite $|M_{127}^{(5)}|$: From the proof of $|M_{63}^{(4)}|$ in [2], the quantity $(1/6)\,\mathrm{tr}(g_{-i}^3)$ is the number of 4-complete subgraphs that contain node i, where g_{-i} is the adjacency matrix formed by the neighbourhood of i. A kite is created by taking one 4-complete subgraph containing i and adding one of the k_i-3 edges that are not in that subgraph, given that $k_i>3$. Hence, there are $(1/6)\,\mathrm{tr}(g_{-i}^3)(k_i-3)$ kites containing node i, and result (10) follows.

(k) ufo $|M_{222}^{(5)}|$: The method of proof is similar to that used for the count of the diamond $|M_{31}^{(4)}|$ in [2]. We can think of a ufo on nodes i and j (both degree 3), and x, y and z (all degree 2) as three distinct paths of length two between i and j, where $(g^2)_{ij}$ represents the number of walks of length 2 between i and j. A ufo is formed by any three of these walks, and so $\binom{(g^2)_{ij}}{3}$ gives the number of distinct ufos that can be built from a pair of degree 3 nodes i and j. Summing across all pairs of nodes i and j will give twice the number of ufos in G, since the edge (i,j) has two endpoints, and we divide the sum by two to give (11).

(l) chevron $|M_{223}^{(5)}|$: The method of proof is similar to that used for the count of the diamond $|M_{31}^{(4)}|$ in [2], and follows immediately from the proof of the ufo count $|M_{222}^{(5)}|$. We can think of a chevron on nodes i and j (both degree 4), and x, y and z (all degree 2), as three distinct triangles with a common edge (i,j). Given this common edge, $(g^2)_{ij}\,(g)_{ij}$ represents the number of walks of length 2 between i and j, that is, the number of distinct triangles in G that contain (i,j). A chevron is formed by any three of these triangles, and so $\binom{(g^2)_{ij}\,(g)_{ij}}{3}$ gives the number of distinct chevrons that can be built from a common edge (i,j), where $k_i>3$ and $k_j>3$. Summing across all pairs of nodes i and j will give twice the number of chevrons in G, since the edge (i,j) has two endpoints, and we divide the sum by two to give (12). See also [5, $n_G(H_{13})$].

(m) hourglass $|M_{235}^{(5)}|$: Let i be the central degree 4 node in an hourglass so that $(1/2)(g^3)_{ii}$ gives the number of triangles that contain node i. An hourglass is formed by any two such triangles, so that $\binom{(1/2)(g^3)_{ii}}{2}$ hourglasses contain node i. We sum over all nodes to give the number of candidate hourglasses in G. The diamond forms an unwanted case, with either of the degree 3 endpoints of the central edge as node i. Hence, we subtract $2\,|M_{31}^{(4)}|$ to give (13). See also [5, $n_G(H_{11})$].

(n) 5-circle $|M_{236}^{(5)}|$: The proof of the 5-circle count proceeds as for the 4-circle count $|M_{30}^{(4)}|$ in [2]. The elements of g^5 are the number of walks of length 5 from node i to node j, and so the trace $\mathrm{tr}(g^5)$ gives the total number of closed walks of length 5 in G. We then prove (14) indirectly. Consider the 5-circle $M_{236}^{(5)}$. There are *ten* ways to traverse the circle (starting at any node and moving clockwise or counterclockwise). However, there are two other ways to walk from a node to itself in 5 steps:

- First, there are ten walks of length 5 through a tadpole $M_{15}^{(4)}$, two from the degree 1 node in the tadpole, four from the degree 3 node, and two from each of the degree 2 nodes.
- Second, there are thirty walks of length 5 around a triangle $M_7^{(3)}$, five from a node to itself, starting in either the clockwise or counterclockwise direction, for each of the three nodes.

Hence, we can write $\text{tr}(g^5) = 10\,|M_{236}^{(5)}| + 10\,|M_{15}^{(4)}| + 30\,|M_7^{(3)}|$, and result (14) follows. See also [15, Theorem 2] and [5, $n_G(C_5)$].

(o) house $|M_{237}^{(5)}|$: Let i and j be the degree 3 nodes in a house, that is formed by one walk of each of lengths 1, 2 and 3 between i and j (there are $(g^3)_{ij}$ walks of length 3, and $(g^2)_{ij}$ walks of length 2). Hence, there are $(1/2)\sum_{i,j:i\neq j}(g^3)_{ij}(g^2)_{ij}(g)_{ij}$ candidate houses in G, where division by two accounts for the two endpoints of edge (i,j). There are three unwanted cases. First, let (i,j) be one of the edges in a triangle. There is one path of lengths 1 or 2, and three walks of length 3, from i to j, giving a total of $9\,|M_7^{(3)}|$ triangles. Second, let i and j be degree 2 and degree 3 nodes in a tadpole (there are two such pairs of nodes). There is one path of lengths 1 or 2, and one walk of length 3 between i and j that is not in the triangle on (i,j), giving $2\,|M_{15}^{(4)}|$ tadpoles. Third, let i and j be a degree 2 node and a degree 3 node in a diamond (there are four such edges). There is one path of lengths 1 or 2, and one walk of length 3 from i to j that is not in the tadpole on (i,j), giving $4\,|M_{31}^{(4)}|$ diamonds. Hence, we subtract $4\,|M_{31}^{(4)}| + 2\,|M_{15}^{(4)}| + 9\,|M_7^{(3)}|$ to give result (15). See also [5, $n_G(H_{12})$].

(p) crown $|M_{239}^{(5)}|$: Consider a 4-path subgraph $M_{13}^{(4)}$ comprised of nodes j, ℓ, r and s. Let each node be in the neighbourhood $\Gamma_G(i)$ of some node i such that $i \neq j \neq \ell \neq r \neq s$. Hence, the five nodes i, j, ℓ, r and s, and the edges between them, form a crown $M_{239}^{(5)}$. The quantity $|M_{13}^{(4)}(g_{-i})|$ gives the number of 4-path subgraphs that are in the neighbourhood of node i, where g_{-i} is the adjacency matrix corresponding to the subgraph formed by $\Gamma_G(i)$. Summing across all nodes i will give the total count of crowns in the graph (16), which can be simplified further by using $|M_{13}^{(4)}| = \sum_{(i,j)\in E}(k_i-1)(k_j-1) - 3\,|M_7^{(3)}|$ from [2] to count 4-paths.

(q) envelope $|M_{254}^{(5)}|$: Let $(i,j) \in E$ be an edge and let $S(i,j)$ denote the common neighbourhood of nodes i and j. Let r and q be two members of S, each of which forms a triangle with common edge (i,j). An envelope is created by matching the diamond on nodes i,j,r,q with a two step path from r to q, through another node x. So, (i,j) connects the degree 3 nodes in an envelope that are not directly-connected to the degree 2 node. Given (i,j), there are $\sum_{\substack{r,q\in S(i,j)\\ r\neq q, k_r>2, k_q>2}} ((g^2)_{rq} - 2)$ envelopes including (i,j), and (17) follows.

(r) lamp $|M_{255}^{(5)}|$: Consider a tadpole subgraph $M_{15}^{(4)}$ comprised of nodes j, ℓ, r and s. Let each node be in the neighbourhood $\Gamma_G(i)$ of some node i such that $i \neq j \neq \ell \neq r \neq s$. Hence, the five nodes i, j, ℓ, r and s, and the edges between them, form a lamp $M_{255}^{(5)}$. The quantity $|M_{15}^{(4)}(g_{-i})|$ gives the number of tadpole subgraphs in the neighbourhood of node i, where g_{-i} is the adjacency matrix corresponding to the subgraph formed by $\Gamma_G(i)$. Summing across all nodes i will give the twice the total count of lamps, since the two degree 4 nodes in the lamp will lead to double-counting, and so we divide by two to give result (18). The result can be simplified further by using $|M_{15}^{(4)}| = (1/2)\sum_{i:k_i>2}(g^3)_{ii}(k_i-2)$ from [2] to count tadpoles.

(s) arrowhead $|M_{507}^{(5)}|$: Consider a 4-circle subgraph $M_{30}^{(4)}$ comprised of nodes j, ℓ, r and s. Let each node be in the neighbourhood $\Gamma_G(i)$ of some node i such that $i \neq j \neq \ell \neq r \neq s$. Hence, the five nodes i, j, ℓ, r and s, and the edges between them, form an arrowhead $M_{507}^{(5)}$. The quantity $|M_{30}^{(4)}(g_{-i})|$ gives the number of 4-circle subgraphs in the neighbourhood of node i, where g_{-i} is the adjacency matrix corresponding to the subgraph formed by $\Gamma_G(i)$. Summing across all nodes i will give the total count of crowns, and (16) follows directly. The result can be simplified further by using $|M_{30}^{(4)}| = (1/8)(\text{tr}(g^4) - 4\,|M_3^{(3)}| - 2m)$ from [2] to count 4-circles.

(t) cat's cradle $|M_{511}^{(5)}|$: Consider a diamond subgraph $M_{31}^{(4)}$ comprised of nodes j, ℓ, r and s. Let each node be in the neighbourhood $\Gamma_G(i)$ of some node i such that $i \neq j \neq \ell \neq r \neq s$. Hence, the five nodes i, j, ℓ, r and s, and the edges between them, form a cat's cradle $M_{507}^{(5)}$. The quantity $|M_{31}^{(4)}(g_{-i})|$ gives the number of diamond subgraphs in the neighbourhood of node i, where g_{-i} is the adjacency matrix corresponding to the subgraph formed by $\Gamma_G(i)$. Summing across all nodes i will give three times the total count of cat's cradles, since each of the degree 4 nodes will count the same cat's cradle, and we divide by three to give result (20). The result can be simplified further by using $|M_{31}^{(4)}| = (1/2)\sum_{i,j:i\neq j}\binom{(g^2)_{ij}(g)_{ij}}{2}$ from [2] to count diamonds.

(u) 5-complete $|M_{1023}^{(5)}|$: See [24, Proposition A.1, eqn. 21].

□

Remark 2.2 (Results, proof strategies and extensions)**.** We make some observations on the techniques that we used to derive Theorem 2.1, and on alternative approaches to counting a given graphlet, and give several generalizations of these methods to the exact counting of larger graphlets.

(i) There are four common strategies that are used in the proofs of (1)–(21):

(a) [Incident structures] Consider a node i or an edge (i, j). Then choose one or more structures that are incident to i or (i, j), such as nodes and/or triangles (5-star $M_{75}^{(5)}$, 5-arrow $M_{77}^{(5)}$, cricket $M_{79}^{(5)}$, bull $M_{87}^{(5)}$, lollipop $M_{117}^{(5)}$ and hourglass $M_{235}^{(5)}$) or the 4-complete subgraph (kite $M_{127}^{(5)}$).

(b) [Walks] Consider nodes i and j, that are not necessarily directly-connected, and examine walks between them, possibly with some additional structure that is incident to one or both of the nodes: for example, the 5-path $M_{86}^{(5)}$, banner $M_{94}^{(5)}$, stingray $M_{95}^{(5)}$, ufo $M_{222}^{(5)}$, chevron $M_{223}^{(5)}$ and house $M_{237}^{(5)}$. If $i = j$ then we obtain a cycle (e.g. the 5-circle $M_{236}^{(5)}$).

(c) [Common neighbourhood] Consider nodes i and j and their common neighbourhood $S(i, j)$. Then look at structure that is incident to *members* of $S(i, j)$, such as the spinning top $M_{119}^{(5)}$ and the envelope $M_{254}^{(5)}$.

(d) [Neighbourhood subgraphs] Consider a node i. Then look at smaller subgraphs that occur in the subgraph formed by the neighbourhood of node i: for example, the crown $M_{239}^{(5)}$ (4-path), lamp $M_{255}^{(5)}$ (tadpole), arrowhead $M_{507}^{(5)}$ (4-circle), cat's cradle $M_{511}^{(5)}$ (diamond), and 5-complete $M_{1023}^{(5)}$ (4-complete subgraph).

We then sum over nodes or edges to obtain the full graphlet count in G, correcting for multiple counting of the same graphlet (e.g. the 5-complete $M_{1023}^{(5)}$), and removing "unwanted cases" that are usually smaller subgraphs which satisfy the same restrictions as the graphlet of interest. For example, a candidate 5-arrow $M_{77}^{(5)}$ is defined as an edge (i, j) combined with two neighbours of i and one neighbour of j; the tadpole graphlet $M_{15}^{(4)}$ satisfies the same restrictions and must be removed (twice). Table 3 displays, for each count formula (1)–(21), the smaller graphlets and "unwanted cases" that explicitly appear in each formula.

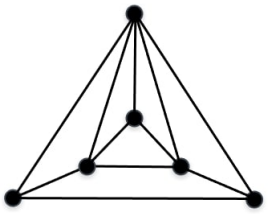
(ii) Observe from (i)d that we could also count the chevron by looking for 4-stars in the neighbourhood of node i. This would give an alternative formulation to (12), namely

$$|M_{223}^{(5)}| = \frac{1}{2}\sum_{i:k_i>3}\sum_{\substack{r:k_r>3\\ r\in V(\Gamma_G(i))}}\binom{k_r}{3} = \frac{1}{12}\sum_{i:k_i>3}\sum_{\substack{r:k_r>3\\ r\in V(\Gamma_G(i))}} k_r(k_r-1)(k_r-2). \tag{22}$$

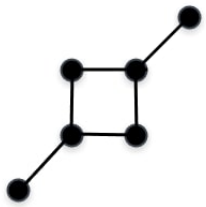
We have found some preliminary evidence in simulations that the chevron can be counted faster on moderately-sized sparse Erdős-Rényi graphs if (22) is used rather than result (12).

	$\lvert M_3^{(3)}\rvert$	$\lvert M_7^{(3)}\rvert$	$\lvert M_{11}^{(4)}\rvert$	$\lvert M_{13}^{(4)}\rvert$	$\lvert M_{15}^{(4)}\rvert$	$\lvert M_{30}^{(4)}\rvert$	$\lvert M_{31}^{(4)}\rvert$	$\lvert M_{63}^{(4)}\rvert$
5-star $\lvert M_{75}^{(5)}\rvert$	·	·	·	·	·	·	·	·
5-arrow $\lvert M_{77}^{(5)}\rvert$	·	·	·	·	X	·	·	·
cricket $\lvert M_{79}^{(5)}\rvert$	·	·	·	·	·	·	·	·
5-path $\lvert M_{86}^{(5)}\rvert$	X	X	X	X	X	·	·	·
bull $\lvert M_{87}^{(5)}\rvert$	·	·	·	·	·	·	X	·
banner $\lvert M_{94}^{(5)}\rvert$	·	·	·	·	·	·	X	·
stingray $\lvert M_{95}^{(5)}\rvert$	·	·	·	·	·	·	·	·
lollipop $\lvert M_{117}^{(5)}\rvert$	·	X	·	·	X	·	X	·
spinning top $\lvert M_{119}^{(5)}\rvert$	·	·	·	·	·	·	·	X
kite $\lvert M_{127}^{(5)}\rvert$	·	·	·	·	·	·	·	X
ufo $\lvert M_{222}^{(5)}\rvert$	·	·	·	·	·	·	·	·
chevron $\lvert M_{223}^{(5)}\rvert$	·	·	·	·	·	·	·	·
hourglass $\lvert M_{235}^{(5)}\rvert$	·	·	·	·	·	·	X	·
5-circle $\lvert M_{236}^{(5)}\rvert$	·	X	·	·	X	·	·	·
house $\lvert M_{237}^{(5)}\rvert$	·	X	·	·	X	·	X	·
crown $\lvert M_{239}^{(5)}\rvert$	·	·	·	X	·	·	·	·
envelope $\lvert M_{254}^{(5)}\rvert$	·	·	·	·	·	·	·	·
lamp $\lvert M_{255}^{(5)}\rvert$	·	·	·	·	X	·	·	·
arrowhead $\lvert M_{507}^{(5)}\rvert$	·	·	·	·	·	X	·	·
cat's cradle $\lvert M_{511}^{(5)}\rvert$	·	·	·	·	·	·	X	·
5-complete $\lvert M_{1023}^{(5)}\rvert$	·	·	·	·	·	·	·	X

Table 3: *This figure represents, for each graphlet count formula in (1)–(21), the smaller graphlets that enter into each result. The graphlets on five nodes, and on three and four nodes, are displayed in Table 1 and Table 2 respectively.*



(a) *Graphlet* $M_{7919}^{(6)}$.



(b) *Graphlet* $M_{1182}^{(6)}$.

Figure 1: *Two six-node graphlets that are used to illustrate extensions of the methods of proof of Theorem 2.1.*

(iii) It is important to apply the strategies in (i) quite carefully to avoid complications that can arise from unwanted cases. Consider the spinning top $M_{119}^{(5)}$ and let (i, j) be the edge between the two degree 3 nodes that are not directly-connected to the degree 1 node. It might seem reasonable to exploit walks of lengths 1, 2 and 4 to count candidate spinning tops by using the formula

$$\frac{1}{2}\sum_{i,j:i\neq j}\binom{(g^2)_{ij}(g)_{ij}}{2}(g^4)_{ij}=\sum_{(i,j)\in E}\binom{(g^2)_{ij}}{2}(g^4)_{ij}.$$

However, the unwanted cases will include five-node graphlets (stingray, crown, chevron, and envelope) as well as *larger* graphlets with six or seven nodes, and this will make the approach unworkable.

(iv) The strategies in (i) can be used to derive exact count formulae for larger graphlets, and we give two examples.

(a) Applying the neighbourhood method based on the house $M_{237}^{(5)}$, we can count the six node graphlet in Figure 1a using the formula

$$|M_{7919}^{(6)}|=\sum_{i:k_i>4}|M_{237}^{(5)}(g_{-i})|.$$

(b) Extending ideas from the banner $M_{94}^{(5)}$, we can count the six node graphlet in Figure 1b using the formula

$$|M_{1182}^{(6)}|=\frac{1}{2}\sum_{\substack{i,j:i\neq j\\ k_i>2,k_j>2}}\binom{(g^2)_{ij}}{2}(k_i-2)(k_j-2)-|M_{31}^{(4)}|-|M_{95}^{(5)}|-3\,|M_{222}^{(5)}|.$$

(v) Several authors present exact enumeration formulae for subgraphs on five nodes, or strategies for deriving combinatorial results. Partial results on five-node subgraphs in the literature, with proof, include [5] (8 subgraphs), [15] (1 subgraph), and [27] (1 subgraph). The eight formulae for subgraphs on five nodes that are listed, with strategies for proof, in [13, §4.4] and [14, Chapter 13], and three of the four subgraphs on five nodes (not including the 5-path) that are given in [4], without proof, all appear in [5], essentially in the same form (although [4, 13] use matrix notation). To our knowledge, the only other complete set of five node results is by [30], although they use a very different method of proof and, in some cases, obtain quite different formulations. We discuss some related issues in Example 3.5, including the 5-path count that is given in [4].

3 Examples

In this section, we present a series of illustrative theoretical and empirical applications of the main result. First, we show how Theorem 2.1 can be used to derive non-trivial analytical results by specializing the count formulae for the 5-path, the bull, and the spinning top subgraphs, to three regular graphs that have been widely studied: the complete graph K_n, the complete N-partite graph $K_{n_1,n_2,\ldots,n_N}$, and the regular ring lattice (also known in the complex systems literature as the small-world model with no edge rewiring [33]). Second, we consider three different formulations of the 5-path count formula from the literature, and show how easily misunderstanding can arise. Third, we show that the formulae in Theorem 2.1 are suitable for empirical application, at least on graphs with a small to moderate number of nodes. We start by proving a straightforward but useful result that gives the number of paths of length k between any pair of nodes in a complete graph K_n, as a function of k and n.

Lemma 3.1 (Number of paths of length k in a complete graph)**.** *Let G be a complete graph K_n with adjacency matrix $g = (g)_{ij}$ and $n \geq 2$. We write $a_k = (g^k)_{ii}$ and $b_k = (g^k)_{ij}$ for the individual elements of g^k. It follows that*

$$b_k = \frac{(n-1)^k + (-1)^{k+1}}{n}; \quad a_k = b_k + (-1)^k.$$

Proof. Observe that the off-diagonal elements of g^k follow the recursion

$$b_k = (n-2)\, b_{k-1} + (n-1)\, b_{k-2}, \quad b_1 = 1, \quad b_2 = (n-2).$$

This is a linear homogeneous equation of order 2 with constant coefficients and characteristic equation

$$P(x) = x^2 - (n-2)\, x - (n-1) = 0.$$

The polynomial $P(x)$ factors as $P(x) = (x+1)(x-n+1)$, and has roots $r_1 = -1$ with multiplicity $m_1 = 1$ and $r_2 = n-1$ with multiplicity $m_2 = 1$. Let C and D be arbitrary polynomials of degree $(m_1 - 1)$ and $(m_2 - 1)$ respectively. The general solution to the recursion is given by

$$b_k = C\,(-1)^k + D\,(n-1)^k.$$

Applying the initial conditions $b_1 = 1$ and $b_2 = (n-2)$, we conclude that

$$(n-1)\, D - C = 1; \quad (n-1)^2\, D + C = n-2.$$

This is a system of linear equations with the unique solution

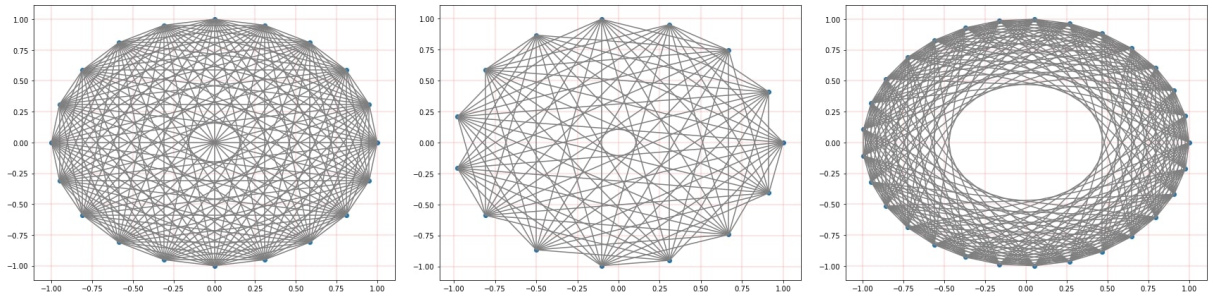
$$C = -1/n; \quad D = 1/n,$$

and the claim on b_k follows immediately. Since $a_k = (n-1)\, b_{k-1}$, one can easily verify that $a_k - b_k = (-1)^k$. □

As a sample application of Lemma 3.1, we specialize it to paths of length 4 in K_n. Directly, we have $(g^4)_{ii} = (n-1)(n^2-3n+3)$ and $g^4_{ij} = (n-2)(n^2-2n+2)$. For example, in K_{20}, illustrated in Figure 2a, there are 6,517 paths of length 4 from any node to itself, and 6,516 paths of length 4 from any node to any other node.

Example 3.2 (Counting 5-paths in the complete graph K_n)**.** We compute each term of the 5-path count formula

$$|M^{(5)}_{86}| = \frac{1}{2} \sum_{i,j:i\neq j} (g^4)_{ij} - 2\,|M^{(3)}_3| - 9\,|M^{(3)}_7| - 3\,|M^{(4)}_{11}| - 2\,|M^{(4)}_{13}| - 2\,|M^{(4)}_{15}|$$



(a) *Complete graph* K_{20}. **(b)** *Complete N-partite,* $(N, n_a) = (5, 3)$. **(c)** *Regular ring lattice,* $(n, k) = (29, 10)$.

Figure 2: *Three regular graphs that are used to illustrate analytic subgraph count formulae in Examples 3.2–3.4.*

given in (4), noting that each node i has degree $k_i = n-1$. From Lemma 3.1 we can write the first term as $(1/2)\sum_{i,j:i\neq j}(g^4)_{ij} = (n-2)(n^2-2n+2)\binom{n}{2}$. The number of 3-stars follows from [2, eqn. 1] as $|M_3^{(3)}| = \sum_i \binom{k_i}{2} = n\binom{n-1}{2}$. Applying Lemma 3.1 and [2, eqn. 2], the number of triangles is $|M_7^{(3)}| = (1/6)\sum_i (g^3)_{ii} = \binom{n}{3}$. From [2, eqn. 3] the number of 4-stars is $|M_{11}^{(4)}| = \sum_i \binom{k_i}{3} = n\binom{n-1}{3}$. Combining [2, eqn. 4] with the triangle count above, the number of 4-paths is $|M_{13}^{(4)}| = \sum_{(i,j)\in E}(k_i-1)(k_j-1) - 3|M_7^{(3)}| = (n-2)^2\binom{n}{2} - 3\binom{n}{3}$. Applying the recurrence $\binom{n}{k} = ((n-k+1)/k)\binom{n}{k-1}$ gives $|M_{13}^{(4)}| = (n-2)(n-3)\binom{n}{2}$. Applying Lemma 3.1 and [2, eqn. 5] gives $|M_{15}^{(4)}| = (1/2)\sum_i (g^3)_{ii}(k_i-2) = 12\binom{n}{4}$ tadpoles. Using the above binomial recurrence twice we have $|M_{15}^{(4)}| = (n-2)(n-3)\binom{n}{2}$ and so $|M_{13}^{(4)}| = |M_{15}^{(4)}|$ in K_n (there is the same number of 4-paths and tadpoles in the complete graph). Putting all this together, and with repeated application of the above recurrence and $\binom{n}{k} = (n/k)\binom{n-1}{k-1}$, we can easily verify that $|M_{86}^{(5)}| = 60\binom{n}{5}$. The combinatorial intuition behind this result is very clear: there are $\binom{n}{5}$ ways to choose 5 nodes, and 60 possible 5-paths between those nodes in the complete graph.

Example 3.3 (Counting bulls in the complete N-partite graph)**.** Let G be a complete N-partite graph $K_{n_1,n_2,\ldots,n_N}$, with nodes partitioned into $N \geq 2$ disjoint sets (or *groups*) such that no pair of nodes within the same group is adjacent, but all other pairs of nodes in the graph are adjacent. Let $n := \sum_c n_c$ denote the number of nodes in the graph, and n_c is the number of nodes in group c. The graph G has an $n \times n$ block-symmetric adjacency matrix g with block $\{a,b\}$ equal to $0_{n_a \times n_a}$ if $a = b$, and equal to $J_{a,b} := \iota_{n_a}\iota'_{n_b}$ if $a \neq b$, for $a, b = 1, 2, \ldots, N$, and $\iota_\cdot$ is a vector of ones. Since $J_{a,b}J_{b,c} = n_b J_{a,c}$, it follows that every element of the block $\{a,b\}$ of g^k will have the same value. It is convenient then to consider blocks rather than individual nodes, and so we define $A^{(1)} := \iota_N\iota'_N - I_N$ and $D := \mathrm{diag}(n_1, n_2, \ldots, n_N)$. It is straightforward to verify that each element of the block $\{a,b\}$ of g^k equals the element $(A^{(k)})_{ab}$, and that $A^{(k)} = A^{(1)}DA^{(k-1)} = (A^{(1)}D)^{k-1}A^{(1)}$. We compute the bull subgraph count formula

$$|M_{87}^{(5)}| = \sum_{\substack{(i,j)\in E \\ k_i>2, k_j>2}} (g^2)_{ij}(k_i-2)(k_j-2) - 2|M_{31}^{(4)}|$$

given in (5), and use the diamond count $|M_{31}^{(4)}| = (1/2)\sum_{(i,j)^\star\in E}(g^2)_{ij}((g^2)_{ij}-1)$ from [2, eqn. 7]. To simplify the example, we assume that every node in a given block c has degree $n - n_c > 2$, so that every edge in the graph is potentially the "central" edge of a bull subgraph. Observe that $(A^{(2)})_{ab} = n - n_a - n_b$ when $a \neq b$. The desired result $|M_{87}^{(5)}|$ may then be rewritten in terms of blocks as

$$|M_{87}^{(5)}| = \sum_{\substack{\{a,b\} \\ a<b}} n_a n_b (n-n_a-n_b)[(n-n_a-2)(n-n_b-2) - (n-n_a-n_b-1)].$$

To go further, we now make the strong assumption that $n_a = n_b$ for all a, b, so that $n = Nn_a$ (every group has the same number of nodes). This allows us to write

$$|M_{87}^{(5)}| = 3\binom{N}{3} n_a^3 \{((N-1)n_a - 2)((N-1)n_a - 3) + n_a - 1\}. \tag{23}$$

Thus we have $|M_{87}^{(5)}| = O(N^5)$ and $|M_{87}^{(5)}| = O(n_a^5)$. The number of bull subgraphs increases as the fifth power of either the number of groups or the number of nodes in each group. Trivially, $|M_{87}^{(5)}| = 0$ when $N \leq 2$ or $n = Nn_a < 5$. Solving (23) for its roots over the positive integers, with $N \geq 2$, gives $(N, n_a) = (3,1), (4,1)$, and it follows that bull subgraphs will be found in all N-partite graphs, with at least three groups, that satisfy the assumptions of this example.

We now consider the combinatorial intuition behind the result (23). Let us choose three groups from N, each of which will contain one of the nodes in a triangle. Given these groups, let us take one of the n_a^3 possible triangles. We then choose one of the three possible pairs of nodes in that triangle to be the degree 3 nodes in a bull subgraph (these nodes, which we denote by i and j, are incident to the "horns" of the bull). There are $3\binom{N}{3} n_a^3$ ways to choose such a triangle and pair of nodes. Let A denote the group that contains node j. There are $((N-1)n_a - 2)$ ways to choose a node k that is adjacent to node i to form the first horn, including the available nodes in group A. If node j is *not* in group A then, given the first horn, there are $((N-1)n_a - 3)$ ways to choose a node ℓ that is adjacent to node j to form the second horn. However, if node j is in group A then there are $n_a - 1$ additional ways to form the second horn (one for each of the possible choices of node j in group A). For example, in the complete 5-partite graph with $n_a = 3$ in each group, illustrated in Figure 2b, there are 74,520 bull subgraphs.

Example 3.4 (Counting spinning tops in the regular ring lattice)**.** Let G be a regular ring lattice on n nodes, where each node is adjacent to its k nearest neighbours in both the clockwise and anti-clockwise directions, giving a total of nk edges, and we require $n > 2k$. We compute the spinning top subgraph formula

$$|M_{119}^{(5)}| = \sum_{\substack{(i,j)\in E \\ k_i>2, k_j>2}} ((g^2)_{ij} - 1) \sum_{\substack{r\in S(i,j) \\ k_r>1}} (k_r - 2) - 12\,|M_{63}^{(4)}|$$

given in (9), where $|M_{63}^{(4)}|$ is the 4-complete subgraph count, and $S(i,j)$ is the set of nodes that are in the joint neighbourhood of nodes i and j. We identify two cases that correspond to a discrete change in the behaviour of the spinning top subgraph count at the threshold $n = 3k+1$.

Case A: $n \geq 3k+1$: We start by considering n sufficiently large relative to k so that we only need to focus on structure among a node's k nearest neighbours in the clockwise direction. Each node is adjacent to $\binom{k}{3}$ 4-complete subgraphs among its k nearest neighbours in the clockwise direction and hence $|M_{63}^{(4)}| = n\binom{k}{3}$. Observe that the degree of a node r in a joint neighbourhood is given by $k_r = 2k$. Then, since g and g^2 are both symmetric Toeplitz, it suffices to find the first row of g^2, and to count the number of elements of the neighbourhood $S(i,j)$, for each edge (i,j), where node j is the a-nearest neighbour of i in the clockwise direction. We exploit these properties to show

$$|M_{119}^{(5)}| = 2\,(k-1)\,n \sum_{a=1}^{k} [2k - (a+1)][2k - (a+2)] - 12n\binom{k}{3}$$

which, after some manipulation, gives the result

$$|M_{119}^{(5)}| = \frac{2}{3} n\,k\,(k-1)\,[7(k-1)(k-2) + 3]. \tag{24}$$

We remark that $|M_{119}^{(5)}| = O(k^4)$ and $|M_{119}^{(5)}| = O(n)$. The number of spinning top subgraphs increases as the fourth power of the degree of each node in the graph, and increases linearly in the number of nodes, for $n \geq 3k+1$.

Case B: $2k+1 \leq n \leq 3k$: As n decreases from $n = 3k$ to $n = 2k+1$, there will be progressively more 4-cliques incident to a given node i, more paths of length 2 between the endpoints of each edge (i, j), and more members of the neighbourhood $S(i, j)$. One can verify that

$$|M_{119}^{(5)}| = 2\,(k-1)\,n \sum_{a=1}^{k} [2k-(a+2)+A][2k-(a+1)+A] - 12\left[n\binom{k}{3}+B\right],$$

where $A = A(a,n,k) = (a-(n-2k-1)\vee 0)$ and $B = B(n,k) = n\,Te_{3k-n}$, where $Te_\ell = \binom{\ell+2}{3}$ denotes the ℓ-th tetrahedral number. Separating out the Case A subgraph count given in (24), we have

$$|M_{119}^{(5)}| = \frac{2}{3}\,nk\,(k-1)\,[7(k-1)(k-2)+3] + 2\,(k-1)\,n \sum_{a=1}^{k} [A(4k-2a-3)+A^2] - 12\,B.$$

Because $A = 0$ as $a \leq n-2k-1$, we have

$$\begin{aligned}
|M_{119}^{(5)}| &= \frac{2}{3}\,nk\,(k-1)\,[7(k-1)(k-2)+3] \\
&\quad + 2\,(k-1)\,n \sum_{a=n-2k-1}^{k} [(a-(n-2k-1))(4k-2a-3)+(a-(n-2k-1))^2] - 12\,n\binom{3k-n+2}{3} \\
&= \frac{2}{3}\,nk\,(k-1)\,[7(k-1)(k-2)+3] + \frac{2}{3}\,n\,(3k-n+1)(3k-n+2)(9k^2-(2n+21)\,k+5n+3) \\
&= \frac{2}{3}\,n\,\{(3k-n+1)(3k-n+2)[9k^2-(2n+21)\,k+5n+3] + [7(k-1)(k-2)+3]\,(k-1)\,k\}.
\end{aligned}$$

Let $f(n)$ denote the subgraph count $|M_{119}^{(5)}|$ for all $n \geq 2k+1$, conditional on k. We now show that $f(n)$ is not monotonically increasing in n. When $k = 1$, there are no spinning tops for any n, and so we can focus on $k \geq 2$. We begin by considering the subgraph count at the boundary between Case A and Case B, and temporarily allow k to take non-integer values. Thus $f(3k+1) - f(3k) = (2/3)\,k\,(k-1)\,(7k^2-39k+35) > 0$ as $k > (\sqrt{541}+39)/14 \approx 4.45$. Since $f(n)$ increases linearly for $n \geq 3k+1$, we conclude that $f(n)$ is not monotonically increasing in n for $k = 2, 3, 4$. This leaves $k \geq 5$. It suffices to show that $f(2k+1) - f(3k) = 2\,k\,(k-1)\,(k^3+9k^2-25k+9) = 2\,k\,(k-1)\,(k\,(k^2+9k-25)+9) > 0$ for all $k \geq 5$. The claim follows because $(k^2+9k-25) > 0$ as $k > (\sqrt{181}-9)/2 \approx 2.23$, while $k = 2$ gives $f(5) - f(6) > 0$. Putting all this together, we see that $|M_{119}^{(5)}|$ is not monotonically increasing in n for any $k \geq 2$, and the number of spinning tops in a regular ring lattice is minimized for some $n \leq 3k+1$. For example, in the regular ring lattice with $k = 10$, we have

$$|M_{119}^{(5)}| = \begin{cases} 488724\,n - 39026\,n^2 + 1092\,n^3 - 10\,n^4, & \text{for } 21 \leq n \leq 30 \\ 30420\,n, & \text{for } n \geq 31 \end{cases}$$

and the number of spinning tops is minimized (with a count of 912,108) for $n = 29$, which is illustrated in Figure 2c.

Example 3.5 (Alternative formulations of the 5-path count formula)**.** We now examine three different formulations of the 5-path count formula that we found in the applied mathematics literature. For reference, we derived result (4):

$$|M_{86}^{(5)}| = \frac{1}{2} \sum_{i,j:i\neq j} (g^4)_{ij} - 2\,|M_3^{(3)}| - 9\,|M_7^{(3)}| - 3\,|M_{11}^{(4)}| - 2\,|M_{13}^{(4)}| - 2\,|M_{15}^{(4)}|.$$

In each case, we begin by stating the authors' results in inverted commas, using their original notation (without defining the terms) before re-stating each result using the notation of our paper. We hope to illustrate the importance of a consistent and clear notation for subgraphs. For an interesting discussion of the importance of good notation, with reference to the field of econometrics, see [1].

Formulation 1 (this is correct): We start with [27, Theorem 4.1], who consider individual matrix elements,

$$\text{“}A = \sum_{i\neq j}[a_{ij}^{(4)} - 2\,a_{ij}^{(2)}\,(d_j - a_{ij})] - \sum_{i=1}^{n}[(2\,d_i - 1)\,a_{ii}^{(3)} + 6\binom{d_i}{3}]\text{”},$$

where A is our notation and refers to *twice* the 5-path count. It follows that

$$\begin{aligned} A/2 &= \frac{1}{2}\left(\sum_{i,j:i\neq j}[(g^4)_{ij} - 2\,(g^2)_{ij}(k_j - (g)_{ij})] - \sum_i[(2k_i-1)(g^3)_{ii} + 6\binom{k_i}{3}]\right) \\ &= \frac{1}{2}\sum_{i,j:i\neq j}(g^4)_{ij} - \sum_{i,j:i\neq j}(g^2)_{ij}(k_j - (g)_{ij}) - \sum_i k_i\,(g^3)_{ii} + \frac{1}{2}\sum_i (g^3)_{ii} - 3\sum_i\binom{k_i}{3}. \end{aligned}$$

We use the following results: from [2, eqn. 3], $-3\sum_i\binom{k_i}{3} = -3\,|M_{11}^{(4)}|$; from [2, eqn. 2], $(1/2)\sum_i (g^3)_{ii} = (1/2)\,\mathrm{tr}(g^3) = 3\,|M_7^{(3)}|$; from [2, eqn. 5], $-\sum_i k_i\,(g^3)_{ii} = -2\,|M_{15}^{(4)}| - 12\,|M_7^{(3)}|$. Hence, $A/2 = |M_{86}^{(5)}|$ as

$$\sum_{i,j:i\neq j}(g^2)_{ij}(k_j - (g)_{ij}) = 2\,(|M_3^{(3)}| + |M_{13}^{(4)}|).$$

Note that $\sum_{i,j:i\neq j}(g^2)_{ij}(k_j - (g)_{ij}) = \sum_{i,j:i\neq j}(g^2)_{ij}(k_j - (g)_{ij} - 1 + 1) = \sum_{i,j:i\neq j}(g^2)_{ij} + \sum_{i,j:i\neq j}(g^2)_{ij}(k_j - (g)_{ij} - 1)$. We combine this decomposition with [27, §2.1 and Theorem 3.7] to show equivalence of (4) and [27, Theorem 4.1].

Formulation 2 (as stated, this is only correct in a special case): We now examine [4, Lemma 9, eqn. 4.12], who present a result that is in terms of a linear combination of small graphlet counts,

$$\text{“}|P_4| = \frac{1}{2}\,\vec{1}^T A^4 \vec{1} - |P_1| - 4\,|P_2| - 2\,|P_3| - 9\,|C_3| - 4\,|C_4| - 6\,|S_{1,3}| - 4\,|S_{T1S}|\text{”}.$$

In our notation this gives

$$|P_4| = \frac{1}{2}\sum_{i,j}(g^4)_{ij} - m - 4\,|M_3^{(3)}| - 9\,|M_7^{(3)}| - 6\,|M_{11}^{(4)}| - 2\,|M_{13}^{(4)}| - 4\,|M_{15}^{(4)}| - 4\,|M_{30}^{(4)}|.$$

Using [2, eqn. 6], it follows that $|M_{86}^{(5)}| = |P_4|$ if and only if $6\,|M_{11}^{(4)}| + 4\,|M_{15}^{(4)}| = 3\,|M_{11}^{(4)}| + 2\,|M_{15}^{(4)}|$. Since counts are non-negative, (4) and [4, Lemma 9, eqn. 4.12] are equivalent if and only if $|M_{11}^{(4)}| = |M_{15}^{(4)}| = 0$, so that G is either a path or a cycle on n nodes. Presumably this reveals an unfortunate double typo in the expression of $|P_4|$.

Formulation 3 (as stated, this is incorrect in a crucial special case): We end with [30, Theorem 9, N_3], who also give a result in terms of smaller graphlet counts,

$$\text{“}N_3 = \sum_i \sum_{(i,j)\in E}(d(j) - 1) - 4\cdot C_4(G) - 2\cdot TT(G) - 3\cdot T(G)\text{”}.$$

The double summation means that each edge is considered in both directions, and rewriting in our notation gives

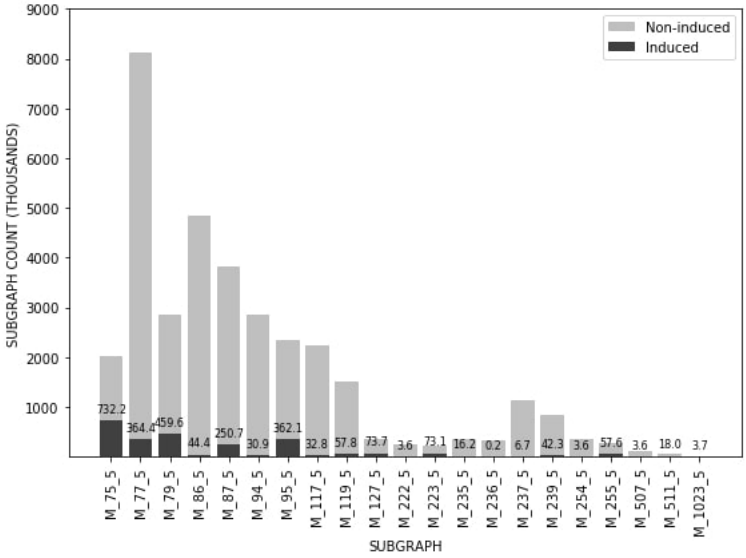
$$N_3 = \sum_{(i,j)\in E} (k_i + k_j - 2) - 3\,|M_7^{(3)}| - 2\,|M_{15}^{(4)}| - 4\,|M_{30}^{(4)}|.$$

We show that this is false by counterexample, taking the count of 5-paths on G, a path of length $n \geq 5$. Immediately, $|M_7^{(3)}| = |M_{11}^{(4)}| = |M_{15}^{(4)}| = 0$ on G, and $|M_3^{(3)}| = (n-2)$ and $|M_{13}^{(4)}| = (n-3)$. Hence, $|M_{86}^{(5)}| = (1/2)\sum_{i,j:i\neq j}(g^4)_{ij} - 4n + 10$ for $n \geq 3$. Observe that $\sum_{i,j:i\neq j}(g^4)_{ij}$ is the sum of non-closed directed walks of length 4. In G, a node can only connect, in four steps, to a node that is two or four steps away. There are 3 ways to do this for each of the two endpoints of the path, and 4 ways to do this for every other node. It follows that $\sum_{i,j:i\neq j}(g^4)_{ij} = 5n - 14$ for $n \geq 5$. Putting all this together gives $|M_{86}^{(5)}| = n - 4$ for $n \geq 5$. To finish, we use $|M_{30}^{(4)}| = 0$ and $N_3 = 2n - 4$ follows. Thus, (4) and [30, Theorem 9, N_3] are not equivalent for any positive number of nodes. For example, when G is a 5-path, we have $|M_{86}^{(5)}| = 1$ as required but $N_3 = 6$.

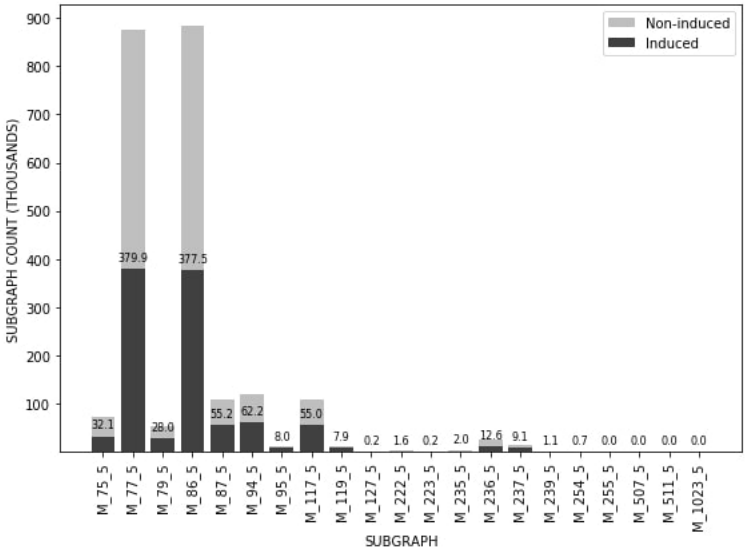
It is sometimes difficult to verify published results on exact subgraph counts, or to distinguish between typos and more substantial errors, and some excellent papers explicitly omit the demonstrations of the count formulae altogether e.g. "The proofs of these results are based on the strategy developed and explained in [14] [our reference] and are not given here as they are lengthy and technical." (Page 267 in [4]) and "It is cumbersome and space-consuming to give proofs of all of these, so we omit them." (Page 1436 in [30]). Important textbook treatments of five-node count formulae mention proof strategies but also omit full and complete proofs e.g. "These calculations are based on the idea of spectral moments analysed in the previous section." (Page 79 in [13]) and "Formulae for a number of simple subgraphs can be derived using very similar techniques to the ones we have encountered so far." (page 136 in [14]). We hope that our main result and its proof will help to fill this gap.

Example 3.6 (An empirical application to real-world network data)**.** We now demonstrate that the analytic formulae of Theorem 2.1 can be implemented practically for empirical applications, at least on small to moderately-sized graphs. It is important to note that we did not attempt to optimize the runtime performance of our algorithms on massive graphs, by using pattern symmetries or memoization, or with lower-level compiled programming languages, or by parallelization and hardware tricks such as use of graphical processing units (GPUs). A serious examination of these computational techniques, and a full speed comparison with other exact and approximate algorithms, would entail a major research programme, and is well beyond the scope of the present paper.[1] Using Python on a Windows machine, we coded separate algorithms for the count of each non-induced five-node subgraph that appears in Table 1. We applied these routines to a small (82 nodes and 522 edges) real-world airline network that is described in detail in [2] (who examine three-node and four-node motifs in airline networks) and [24] (who use analytic five-node subgraph counts on the 5-star, 5-path, 5-arrow, kite, and 5-complete subgraphs to compute generalized clustering coefficients of order five). We computed each individual induced count in Theorem A.1. The counts are displayed in Figure 3. We observe that there is considerable variation in the frequency of occurrence of induced graphlets in this network: the 5-star, cricket, bull and stingray are quite common relative to their occurrence in the Erdős-Rényi random graph, while the 5-path and 5-circle arise much less often. Such observations could potentially be useful in classifying real-world networks, or as a part of edge prediction techniques, as is also noted by [30], as well as for understanding the fundamental structural properties of such networks. In Table 4, we plot representative non-induced subgraphs, with the corresponding identities of each node.

[1] There is a rich and extensive literature in applied mathematics and computer science on the computational aspects of subgraph enumeration, and the design of efficient exact and sampling algorithms. To the best of our knowledge, the most complete treatment of practical five-node subgraph counting using exact methods is the excellent [30], who construct algorithms based on cutting subgraphs into a small set of smaller subgraphs that can then be exhaustively and rapidly enumerated, even for massive graphs. Related papers that discuss graphlet counting algorithms, with some reference to large networks, include [8, 10, 11, 17, 21, 29]; also see the survey paper by [31].



(a) *The subgraph counts for each non-induced (Theorem 2.1) and induced (Theorem A.1) five-node subgraph, from the route network of Southwest Airlines in 2013Q4. The induced counts are given in numbers above the dark bars.*



(b) *The subgraph counts for each non-induced (Theorem 2.1) and induced (Theorem A.1) five-node subgraph, from the Erdős-Rényi random graph $G(n,p)$, with the number of nodes and edge-probability p set equal to the same number of nodes ($n = 88$) and the same density ($p \approx 0.1364$) as the route network of Southwest Airlines in 2013Q4. The induced counts are given in numbers above the dark bars. Note that the scaling of the y-axis is an order of magnitude lower than in Figure 3a above.*

Figure 3: *This figure demonstrates that the 21 exact subgraph count formulae for non-induced and induced graphlets on five nodes can all be conveniently coded for empirical applications, as discussed in Example 3.6.*

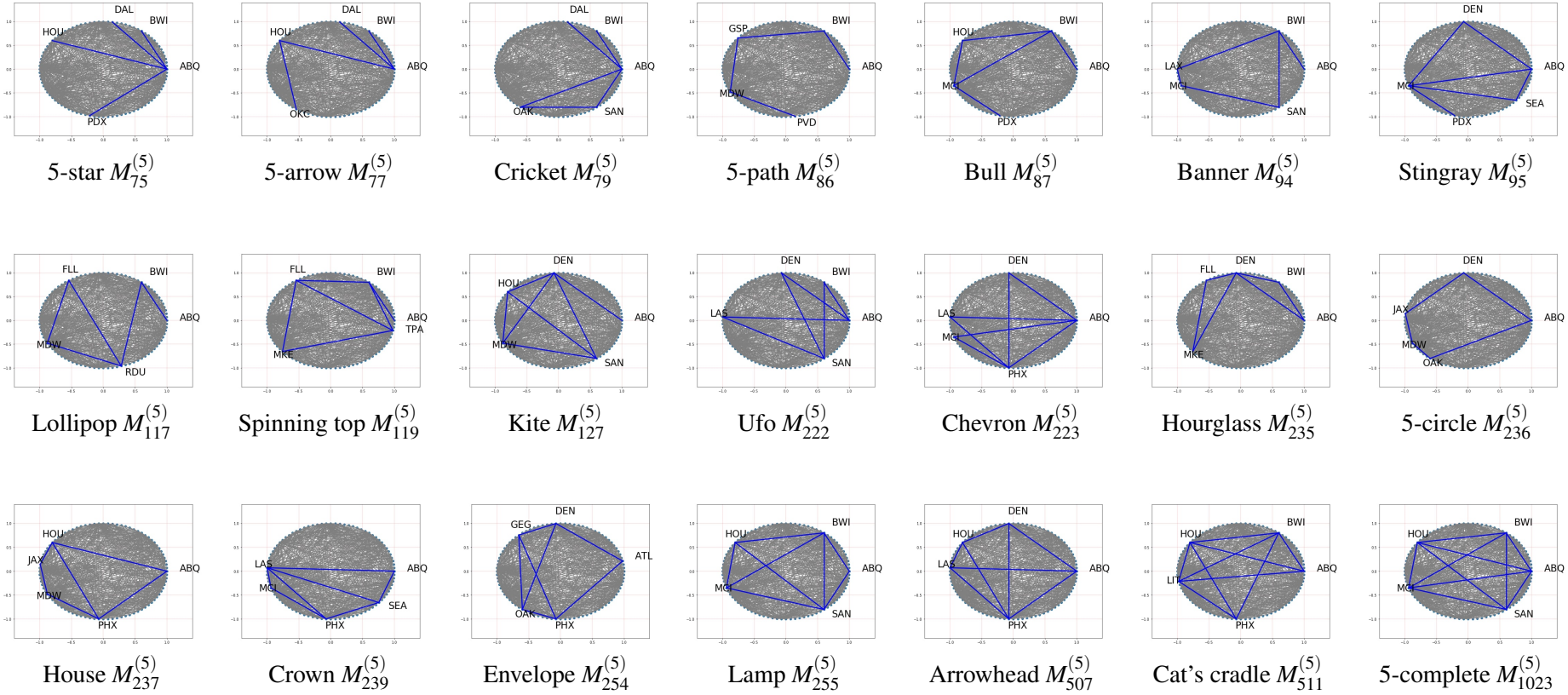


Table 4: *Representative non-induced subgraphs on five nodes, from the real-world route network of Southwest Airlines in 2013Q4, as discussed in Example 3.6. The identity of the airport corresponding to each three-letter IATA code can be found at http://www.iata.org/en/publications/directories/code-search/.*

4 Conclusions

Efficient and elegant subgraph count formulae can lead to a better understanding of the topological structure and related function of real-world networks, and can provide improved insight into widely-used graph statistics such as the assortativity index and the overall clustering coefficient. Much of the focus of research in applied mathematics since the 1970s and, more recently, in theoretical and applied computer science, has been on development of faster implementations of such formulae, and their application to ever larger datasets. This is of deep theoretical interest and crucial empirical importance. However, there has been relatively little work on elementary combinatorial treatments of small subgraph counts. While some of these formulae and methods of proof now appear in important textbooks on graph theory, their correct derivation can be surprisingly tricky, and new formulations of exact count formulae can be conveniently specialized to find non-trivial properties of various theoretical graphs. Here we have presented full results for the 21 induced and non-induced exact graphlet counts on five nodes, some of which appear to be new. We place particular emphasis on the combinatorial intuition behind the results and illustrate, through a series of examples, how these formulae can be used in theoretical and empirical work. While exact formulae might, in future, motivate some useful computational improvements, we expect that they will be most useful in deriving theoretical results for special graphs and statistics. It would be very difficult, using current techniques, to derive complete sets of exact graphlet enumeration formulae for more than five nodes (for instance, there are 112 distinct graphlets on six nodes, and 853 distinct graphlets on seven nodes), even before we consider efficient implementation. One can perhaps envisage a role for computer-assisted (or automated) theorem proving in working towards this goal.

Acknowledgements

We are grateful to Jack Lawford for helpful discussions about Table 1, and for finding some isomorphic subgraphs. The usual caveat applies. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Keywords: Graph theory, graphlet, subgraph counting.

PACS numbers: 02.10.Ox (Combinatorics; graph theory), 89.75.-k (Complex systems).

JEL classification: C65 (Miscellaneous Mathematical Tools).

A Counting induced graphlets

It is well-known that induced subgraph counts $\widetilde{y} = (|\widetilde{M}_a^{(b)}|)$ follow from non-induced subgraph counts $y = (|M_a^{(b)}|)$ by a simple linear combination $\widetilde{y} = y - A\widetilde{y}$. Hence, $\widetilde{y} = (I+A)^{-1} y$, where invertibility of $I+A$ follows immediately from the properties of a unit upper-triangular matrix. The matrix A for five-node graphlets is given in Table B.1. We report individual count formulae in Theorem A.1.

Theorem A.1 (Count formulae for induced graphlets on five nodes)**.**

$$
\begin{aligned}
|\widetilde{M}_{75}^{(5)}| &= |M_{75}^{(5)}| - |M_{79}^{(5)}| + |M_{95}^{(5)}| - |M_{127}^{(5)}| - 2\,|M_{223}^{(5)}| + |M_{235}^{(5)}| - |M_{239}^{(5)}| + 2\,|M_{255}^{(5)}| + |M_{507}^{(5)}| - 3\,|M_{511}^{(5)}| + 5\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{77}^{(5)}| &= |M_{77}^{(5)}| - 2\,|M_{79}^{(5)}| - 2\,|M_{87}^{(5)}| - 2\,|M_{94}^{(5)}| + 5\,|M_{95}^{(5)}| - |M_{117}^{(5)}| + 4\,|M_{119}^{(5)}| - 9\,|M_{127}^{(5)}| + 6\,|M_{222}^{(5)}| - 12\,|M_{223}^{(5)}|\\
&\quad + 4\,|M_{235}^{(5)}| + 4\,|M_{237}^{(5)}| - 10\,|M_{239}^{(5)}| - 10\,|M_{254}^{(5)}| + 20\,|M_{255}^{(5)}| + 20\,|M_{507}^{(5)}| - 36\,|M_{511}^{(5)}| + 60\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{79}^{(5)}| &= |M_{79}^{(5)}| - 2\,|M_{95}^{(5)}| + 3\,|M_{127}^{(5)}| + 6\,|M_{223}^{(5)}| - 2\,|M_{235}^{(5)}| + 3\,|M_{239}^{(5)}| - 8\,|M_{255}^{(5)}| - 4\,|M_{507}^{(5)}| + 15\,|M_{511}^{(5)}| - 30\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{86}^{(5)}| &= |M_{86}^{(5)}| - |M_{87}^{(5)}| - 2\,|M_{94}^{(5)}| + 2\,|M_{95}^{(5)}| - 2\,|M_{117}^{(5)}| + 4\,|M_{119}^{(5)}| - 6\,|M_{127}^{(5)}| + 6\,|M_{222}^{(5)}| - 6\,|M_{223}^{(5)}| + 4\,|M_{235}^{(5)}|\\
&\quad - 5\,|M_{236}^{(5)}| + 7\,|M_{237}^{(5)}| - 10\,|M_{239}^{(5)}| - 14\,|M_{254}^{(5)}| + 18\,|M_{255}^{(5)}| + 24\,|M_{507}^{(5)}| - 36\,|M_{511}^{(5)}| + 60\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{87}^{(5)}| &= |M_{87}^{(5)}| - 2\,|M_{95}^{(5)}| - 2\,|M_{119}^{(5)}| + 6\,|M_{127}^{(5)}| + 6\,|M_{223}^{(5)}| - |M_{237}^{(5)}| + 5\,|M_{239}^{(5)}| + 4\,|M_{254}^{(5)}| - 14\,|M_{255}^{(5)}| - 12\,|M_{507}^{(5)}|\\
&\quad + 30\,|M_{511}^{(5)}| - 60\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{94}^{(5)}| &= |M_{94}^{(5)}| - |M_{95}^{(5)}| - |M_{119}^{(5)}| + 3\,|M_{127}^{(5)}| - 6\,|M_{222}^{(5)}| + 6\,|M_{223}^{(5)}| - 2\,|M_{237}^{(5)}| + 4\,|M_{239}^{(5)}| + 8\,|M_{254}^{(5)}| - 12\,|M_{255}^{(5)}|\\
&\quad - 16\,|M_{507}^{(5)}| + 30\,|M_{511}^{(5)}| - 60\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{95}^{(5)}| &= |M_{95}^{(5)}| - 3\,|M_{127}^{(5)}| - 6\,|M_{223}^{(5)}| - 2\,|M_{239}^{(5)}| + 10\,|M_{255}^{(5)}| + 4\,|M_{507}^{(5)}| - 24\,|M_{511}^{(5)}| + 60\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{117}^{(5)}| &= |M_{117}^{(5)}| - 2\,|M_{119}^{(5)}| + 3\,|M_{127}^{(5)}| - 4\,|M_{235}^{(5)}| - 2\,|M_{237}^{(5)}| + 6\,|M_{239}^{(5)}| + 6\,|M_{254}^{(5)}| - 12\,|M_{255}^{(5)}| - 16\,|M_{507}^{(5)}|\\
&\quad + 30\,|M_{511}^{(5)}| - 60\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{119}^{(5)}| &= |M_{119}^{(5)}| - 3\,|M_{127}^{(5)}| - 2\,|M_{239}^{(5)}| - 2\,|M_{254}^{(5)}| + 8\,|M_{255}^{(5)}| + 8\,|M_{507}^{(5)}| - 24\,|M_{511}^{(5)}| + 60\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{127}^{(5)}| &= |M_{127}^{(5)}| - 2\,|M_{255}^{(5)}| + 6\,|M_{511}^{(5)}| - 20\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{222}^{(5)}| &= |M_{222}^{(5)}| - |M_{223}^{(5)}| - |M_{254}^{(5)}| + |M_{255}^{(5)}| + 2\,|M_{507}^{(5)}| - 4\,|M_{511}^{(5)}| + 10\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{223}^{(5)}| &= |M_{223}^{(5)}| - |M_{255}^{(5)}| + 3\,|M_{511}^{(5)}| - 10\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{235}^{(5)}| &= |M_{235}^{(5)}| - |M_{239}^{(5)}| + 2\,|M_{255}^{(5)}| + 2\,|M_{507}^{(5)}| - 6\,|M_{511}^{(5)}| + 15\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{236}^{(5)}| &= |M_{236}^{(5)}| - |M_{237}^{(5)}| + |M_{239}^{(5)}| + 2\,|M_{254}^{(5)}| - 2\,|M_{255}^{(5)}| - 4\,|M_{507}^{(5)}| + 6\,|M_{511}^{(5)}| - 12\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{237}^{(5)}| &= |M_{237}^{(5)}| - 2\,|M_{239}^{(5)}| - 4\,|M_{254}^{(5)}| + 6\,|M_{255}^{(5)}| + 12\,|M_{507}^{(5)}| - 24\,|M_{511}^{(5)}| + 60\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{239}^{(5)}| &= |M_{239}^{(5)}| - 4\,|M_{255}^{(5)}| - 4\,|M_{507}^{(5)}| + 18\,|M_{511}^{(5)}| - 60\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{254}^{(5)}| &= |M_{254}^{(5)}| - |M_{255}^{(5)}| - 4\,|M_{507}^{(5)}| + 9\,|M_{511}^{(5)}| - 30\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{255}^{(5)}| &= |M_{255}^{(5)}| - 6\,|M_{511}^{(5)}| + 30\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{507}^{(5)}| &= |M_{507}^{(5)}| - 3\,|M_{511}^{(5)}| + 15\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{511}^{(5)}| &= |M_{511}^{(5)}| - 10\,|M_{1023}^{(5)}|.\\
|\widetilde{M}_{1023}^{(5)}| &= |M_{1023}^{(5)}|.
\end{aligned}
$$

Proof of Theorem A.1. All of the results follow directly from the inversion of $(I+A)^{-1}$. □

To develop some combinatorial intuition, we consider the envelope count $|\widetilde{M}^{(5)}_{254}|$. Let k–k' denote an edge between a degree k node and any other degree k' node. To build an envelope, there is one way to remove one edge from a lamp (between the pair of degree 4 nodes) and four ways to remove one edge from an arrowhead (between the degree 4 node and any of the degree 3 nodes). Further, there are six ways to remove one edge from a cat's cradle to obtain a lamp (between a degree 3 and a degree 4 node) and three ways to remove one edge from a cat's cradle to obtain an arrowhead (between any of the degree 4 nodes). It follows that there are nine ways to remove two edges from a cat's cradle to obtain an envelope (there are $(6\times 1)+(3\times 4)=18$ ways to remove two edges from a cat's cradle to obtain an envelope, i.e. remove edge 4–4 then edge 3–4, or remove edge 3–4 then edge 4–4; hence, we divide by two to correct for double-counting). Finally, there are thirty ways to remove three edges from a 5-complete to obtain an envelope (remove any of the 10 edges to move to a cat's cradle; the subsequent edge removals are either (a) edges 4–4 and 3–4, or (b) edges 3–4 and 4–4; this amounts to removing any of the 10 edges followed by any two of the three edges in the triangle that is not incident to any of the endpoints of the first edge that was removed; there are $10\times 3=30$ ways to do this). Putting all this together we have the formula

$$|\widetilde{M}^{(5)}_{254}| = |M^{(5)}_{254}| - |\widetilde{M}^{(5)}_{255}| - 4\,|\widetilde{M}^{(5)}_{507}| - 9\,|\widetilde{M}^{(5)}_{511}| - 30\,|M^{(5)}_{1023}|,$$

and the result follows. A full proof of Theorem A.1 using this combinatorial approach would be very cumbersome.

Remark A.2. With regards to the construction of matrix A in Table B.1 we note that:

(i) The induced graphlet counts are of the form $|\widetilde{M}^{(b)}_a| = |M^{(b)}_a| - \sum_{s>a} c_s\,|\widetilde{M}^{(b)}_s|$, where c_s are coefficients from A, i.e. induced count = non-induced count − correction. The correction term cannot include any graphlet $\widetilde{M}^{(b)}_s$ with $s \le a$. This follows from the observation that the subgraphs are ordered by increasing a, which encodes the *lowest* binary form of the graphlet's adjacency matrix (see Section 1.1). An induced subgraph count is a non-induced subgraph count corrected for subgraphs that have more edges, and this requires removal of at least one edge. Hence, the lower-triangle and main diagonal of A are all zero. For example, we cannot remove one edge from a stingray $\widetilde{M}^{(5)}_{95}$ ($m=6$) to give a lollipop $\widetilde{M}^{(5)}_{117}$ ($m=5$) since $95<117$ and removal of any edge (corresponding to a 1 in the adjacency matrix) will necessarily reduce a.

(ii) The graphlet $\widetilde{M}^{(b)}_s$ must have a larger number of edges than $\widetilde{M}^{(b)}_a$. For example, we cannot remove any edges from an hourglass $\widetilde{M}^{(5)}_{235}$ ($m=6$) to give a chevron $\widetilde{M}^{(5)}_{223}$ ($m=7$), even though $s=235>a=223$.

(iii) The degree distribution $P(k)$ uniquely determines the subgraph (although there can be missing information on the actual links between nodes of a given degree) for all five-node graphlets except for the banner and the lollipop (both have degree distribution $P(k)=1,2,2,2,3$) and the ufo and the house (both of which have $P(k)=2,2,2,3,3$). However, $P(k)$ can still give some intuition as to which edges to remove when moving from one subgraph to another. Let $k^{(i)}$ denote the degrees of a graphlet in ascending order. An element of A is zero if the degree distribution of the candidate correction term $\widetilde{M}^{(b)}_s$ has $k^{(i)}$ less than the corresponding $k^{(i)}$ of the target $\widetilde{M}^{(b)}_a$, because the degree of a node cannot be increased by removal of an edge. For example, a banner ($P(k)=1,2,2,2,3$) cannot reduce to a 5-star ($P(k)=1,1,1,1,4$). Two further examples are: an hourglass ($P(k)=2,2,2,2,4$) cannot reduce to a bull ($P(k)=1,1,1,2,3$) and a chevron ($P(k)=2,2,2,4,4$) cannot reduce to a spinning top ($P(k)=1,2,3,3,3$).

(iv) Applying remarks (i)–(iii) leaves only six zero elements of A: (a) a ufo does not reduce to a bull because the degree 2 nodes are not connected to one another so $P(k)=2,2,2,3,3$ does not reduce to $P(k)=1,1,2,3,3$; (b) an hourglass reduces to a lollipop (but not to a banner, which has the same degree distribution as the

5-star $\lvert M_{75}^{(5)}\rvert$	0	0	1	0	0	0	1	0	0	1	0	2	1	0	0	1	0	2	1	3	5
5-arrow $\lvert M_{77}^{(5)}\rvert$	0	0	2	0	2	2	5	1	4	9	6	12	4	0	4	10	10	20	20	36	60
cricket $\lvert M_{79}^{(5)}\rvert$	0	0	0	0	0	0	2	0	0	3	0	6	2	0	0	3	0	8	4	15	30
5-path $\lvert M_{86}^{(5)}\rvert$	0	0	0	0	1	2	2	2	4	6	6	6	4	5	7	10	14	18	24	36	60
bull $\lvert M_{87}^{(5)}\rvert$	0	0	0	0	0	0	2	0	2	6	0	6	0	0	1	5	4	14	12	30	60
banner $\lvert M_{94}^{(5)}\rvert$	0	0	0	0	0	0	1	0	1	3	6	6	0	0	2	4	8	12	16	30	60
stingray $\lvert M_{95}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	3	0	6	0	0	0	2	0	10	4	24	60
lollipop $\lvert M_{117}^{(5)}\rvert$	0	0	0	0	0	0	0	0	2	3	0	0	4	0	2	6	6	12	16	30	60
spinning top $\lvert M_{119}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	2	2	8	8	24	60
kite $\lvert M_{127}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	6	20
ufo $\lvert M_{222}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	2	4	10
chevron $\lvert M_{223}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	10
hourglass $\lvert M_{235}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	2	6	15
5-circle $\lvert M_{236}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2	4	6	12
house $\lvert M_{237}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4	6	12	24	60
crown $\lvert M_{239}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4	18	60
envelope $\lvert M_{254}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	9	30
lamp $\lvert M_{255}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	30
arrowhead $\lvert M_{507}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	15
cat's cradle $\lvert M_{511}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
5-complete $\lvert M_{1023}^{(5)}\rvert$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table B.1: *This figure represents the matrix $A = (A)_{ij}$, and gives the number of copies of each graphlet i in graphlet j, where $i \neq j$. The row and column orderings are identical. The same figure, up to a row and column re-ordering, is given as $I + A$ in Figure 12 of a working paper version of [30] that can be found at http://arxiv.org/abs/1610.09411, where I is the identity matrix. See Theorem A.1 and Remark A.2 for details.*

lollipop); (c) a ufo reduces to a banner (but not to a lollipop, for the same reason); (d) a chevron reduces to a banner (through a stingray or a ufo, but not to a lollipop, as above); (e) an arrowhead reduces to a crown or an envelope but not to a kite; (f) a crown reduces to a house but not to a ufo.

(v) For the remaining non-zero elements of A, it is straightforward to calculate these analytically, or numerically, from (1)–(21). Consider the count of 5-stars $M_{77}^{(5)}$ in a chevron $M_{223}^{(5)}$. A chevron has $P(k) = 2,2,2,4,4$, and the 5-star count (1) immediately gives $\sum_{i:k_i>3} \binom{k_i}{4} = 2$. A slightly more involved example is the count of crickets $M_{79}^{(5)}$ in the 5-complete $M_{1023}^{(5)}$. The 5-complete has $P(k) = 4,4,4,4,4$, and the cricket count (3) gives $(1/2)\sum_{i:k_i>3}(g^3)_{ii}\binom{k_i-2}{2} = (1/2)\,\mathrm{tr}(g^3) = 3\,|M_7^{(3)}|$ from [2, eqn. 2]. Since there are $\binom{n}{3}$ triangles in K_n, there are $3\binom{5}{3} = 30$ crickets in a 5-complete subgraph. Alternatively, each node in K_n can be the center of a 5-star, so there are $n\binom{n-1}{4} = 5\binom{n}{5}$ 5-stars in K_n and five 5-stars in K_5. A cricket is obtained by adding one more edge to a 5-star. There are $\binom{4}{2} = 6$ ways to do this, and so there are $6 \times 5 = 30$ crickets in the 5-complete subgraph.

References

[1] K.M. Abadir and J.R. Magnus. Notation in econometrics: a proposal for a standard. *Econometrics Journal*, 5: 76–90, 2002.

[2] M. Agasse-Duval and S. Lawford. Subgraphs and motifs in a dynamic airline network. Technical Report arXiv:1807.02585, 2018.

[3] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74: 47–97, 2002.

[4] A. Allen-Perkins, J.M. Pastor, and E. Estrada. Two-walks degree assortativity in graphs and networks. *Applied Mathematics and Computation*, 311:262–271, 2017.

[5] N. Alon, R. Yuster, and U. Zwick. Finding and counting given length cycles. *Algorithmica*, 17:209–223, 1997.

[6] U. Alon. Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, 8:450–461, 2007.

[7] A.R. Benson, D.F. Gleich, and J. Leskovec. Higher-order organization of complex networks. *Science*, 353: 163–166, 2016.

[8] S.K. Bera, N. Pashanasangi, and C. Seshadhri. Linear time subgraph counting, graph degeneracy, and the chasm at size six. In *Proceedings of the 11th Innovations in Theoretical Computer Science Conference (ITCS2020)*, pages 38:1–38:20, 2020.

[9] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424:175–308, 2006.

[10] X. Chen and J.C.S. Lui. Mining graphlet counts in online social networks. *ACM Transactions on Knowledge Discovery from Data*, 12:41:1–41:38, 2018.

[11] R. Curticapean, H. Dell, and D. Marx. Homomorphisms are a good basis for counting small subgraphs. In *Proceedings of the 49th Annual ACM SIGACT Symposium of Theory of Computing (STOC 2017)*, pages 210–223, 2017.

[12] E. Estrada. Topological structural classes of complex networks. *Physical Review E*, 75:016103, 2007.

[13] E. Estrada. *The Structure of Complex Networks*. Oxford University Press, 2011.

[14] E. Estrada and P.A. Knight. *A First Course in Network Theory*. Oxford University Press, 2015.

[15] F. Harary and B. Manvel. On the number of cycles in a graph. *Matematický časopis*, 1:55–63, 1971.

[16] W. Hayes, K. Sun, and N. Pržulj. Graphlet-based measures are suitable for biological network comparison. *Bioinformatics*, 34:483–491, 2013.

[17] T. Hočevar and J. Demšar. Computation of graphlet orbits for nodes and edges in sparse graphs. *Journal of Statistical Software*, 71:1–24, 2016.

[18] A. Itai and M. Rodeh. Finding a minimum circuit in a graph. *SIAM Journal on Computing*, 7:413–423, 1978.

[19] S. Itzkovitz and U. Alon. Subgraphs and network motifs in geometric networks. *Physical Review E*, 71:026117, 2005.

[20] M.O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.

[21] D. Kane, K. Mehlhorn, T. Sauerwald, and H. Sun. Counting arbitrary subgraphs in data streams. In *Proceedings of the 39th International Colloquium on Automata, Languages, and Programming (ICALP 2012)*, pages 598–609, 2012.

[22] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Topological generalizations of network motifs. *Physical Review E*, 70:031909, 2004.

[23] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj. Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society Interface*, 7:1341–1354, 2010.

[24] S. Lawford and Y. Mehmeti. Cliques and a new measure of clustering: with application to U.S. domestic airlines. *Physica A*, 560:125158, 2020.

[25] T. Milenković and N. Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, 6:257–273, 2008.

[26] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[27] N. Movarraei and M.M. Shikare. On the number of paths of lengths 3 and 4 in a graph. *International Journal of Applied Mathematical Research*, 3:178–189, 2014.

[28] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[29] N. Pashanasangi and C. Seshadhri. Efficiently counting vertex orbits of all 5-vertex subgraphs, by EVOKE. Technical Report arXiv:1911.10616, 2019.

[30] A. Pinar, C. Seshadhri, and V. Vishal. ESCAPE: Efficiently counting all 5-vertex subgraphs. In *Proceedings of the International World Wide Web Conference (IW3C2)*, pages 1431–1440, 2017.

[31] P. Ribeiro, P. Paredes, M.E.P. Silva, D. Aparício, and F. Silva. A survey on subgraph counting: Concepts, algorithms and applications to network motifs and graphlets. Technical Report arXiv:1910.13011, 2019.

[32] S.H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[33] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[34] F. Xia, H. Wei, S. Yu, D. Zhang, and B. Xu. A survey of measures for network motifs. *IEEE Access*, 7: 106576–106587, 2019.