Random multigraphs and aggregated triads with fixed degrees

DOI:
10.1017/nws.2017.31

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Frank, O., & Shafie, T. (2018). Random multigraphs and aggregated triads with fixed degrees. *Network Science*, *6*(2), 232-250. https://doi.org/10.1017/nws.2017.31

Published in:
Network Science

Random multigraphs and aggregated triads with fixed degrees

OVE FRANK
Department of Statistics, Stockholm University, Sweden
(*e-mail:* `ove.frank@stat.su.se`)

TERMEH SHAFIE
Department of Computer & Information Science, University of Konstanz, Germany
(*e-mail:* `termeh.shafie@uni-konstanz.de`)

Abstract

Random multigraphs with fixed degrees are obtained by the configuration model or by so called random stub matching. New combinatorial results are given for the global probability distribution of edge multiplicities and its marginal local distributions of loops and edges. The number of multigraphs on triads is determined for arbitrary degrees, and aggregated triads are shown to be useful for analyzing regular and almost regular multigraphs. Relationships between entropy and complexity are given and numerically illustrated for multigraphs with different number of vertices and specified average and variance for the degrees.

Keywords: *multigraph, fixed degrees, edge multiplicity, configuration model, random stub matching, complexity and simplicity, entropy, aggregated triad*

1 Introduction and Overview

A random multigraph with n vertices and specified degree sequence $\mathbf{d} = (d_1,\ldots,d_n)$ with $d_1+\ldots+d_n = 2m$ is obtained by giving equal probabilities to all permutations of $2m$ vertex labels (stubs) chosen such that vertex i occurs d_i times for $i = 1,\ldots,n$. The m unordered pairs of stubs in a permutation are interpreted as sites for edges. This random model is called random stub matching (RSM) (Shafie, 2016) and is in combinatorics also called the configuration model (Janson, 2009; Wormald, 1999; McKay & Wormald, 1991; Bender & Canfield, 1978).

The numbers of edges at different sites are given by random multiplicities $\mathbf{M} = (M_{ij} : 1 \leq i \leq j \leq n)$ and its RSM-distribution is derived and specified in Section 2, Theorem 1. The minimal sufficient statistic T for this distribution is shown to be a global complexity measure, and conditional on T the distribution of $\mathbf{M}$ is uniform. Complexity $T = 0$ corresponds to simple graphs, and loops and multiple edges increase complexity. The outcomes of $\mathbf{M}$ have probabilities that decrease with increasing complexity.

In Section 3, the loop distributions at local vertices as well as distributions of loops and edges at local dyads of vertices are investigated under RSM. In particular, Theorem 2 gives a general formula for the probability of an arbitrary number of loops at a vertex, which is obtained as a marginal distribution of M_{ii} in the global distribution of $\mathbf{M}$. Using similar technique, the marginal trivariate distribution of (M_{ii}, M_{ij}, M_{jj}) is specified in Theorem 3. The M_{ij}-distribution can be obtained by marginal summation in the trivariate distribution. A closed expression for the bivariate distribution of (M_{ii}, M_{ij}) is given in Theorem 4, which allows a simpler marginal summation in order to get the M_{ij}-distribution.

A global logarithmic measure of how many outcomes of $\mathbf{M}$ that are essential in a probabilistic sense is the entropy $H(\mathbf{M})$, which is bounded by $\log_2 K(\mathbf{d})$, the logarithm of the total number of outcomes. Usually, all outcomes of $\mathbf{M}$ and their probabilities are needed to calculate the entropy. Section 4 defines entropy and shows that distributions of local multiplicities are sufficient to determine the global entropy $H(\mathbf{M})$. Calculations are simplified by using an algorithm that identifies all aggregated triads with distinct degree sequences. This approach makes it unnecessary to list all the outcomes of $\mathbf{M}$ in order to calculate its entropy. The focus on triads is also important since it facilitates the possibilities to estimate entropy by sampling methods, which are briefly commented on.

The use of triads has a long tradition in social network theory and both undirected, directed, and colored triads for graphs are well known. See, for instance, the pioneering papers by Holland & Leinhardt (1976) and Frank & Strauss (1986), or the textbooks by Wasserman & Faust (1994) and Kolaczyk (2009), or the paper by Frank (1988) on combinatorial counts of general triads. The number of possible multigraphs on triads is given in Theorem 6 for any specified degree sequence. This knowledge is useful for checking and control of the procedures of entropy calculation and the development of sampling methods mentioned above. For a general number of vertices $n > 3$, the number $K(\mathbf{d})$ of multigraphs with degree restrictions seems to be unknown.

Section 5 treats some illustrating examples. For certain specifications of the degree sequence $\mathbf{d}$, it is possible to determine the entropy $H(\mathbf{M})$ and the complexity T with particularly simple versions of the methods. The simplest case is when all aggregated triads have the same degree sequence, which holds true only for regular multigraphs with all degrees equal. For regular multigraphs, the global entropy and expected complexity are determined as functions of the common local distribution of edges at any aggregated triad. For almost regular graphs, with specified average degree and small degree variance, exact and approximate formulae are obtained and numerically illustrated.

2 Random Multigraph Model

Consider the site space $R = \{(i, j) : 1 \leq i \leq j \leq n\}$ for edges, where vertex pair (i, j) is a canonical representation for an undirected edge between vertices i and j in $V = \{1, \ldots, n\}$. The multigraph model RSM($\mathbf{d}$) is defined as an undirected multigraph with m edges obtained by random stub matching of $2m$ stubs of which d_i are attached to vertex $i \in V$ and $\sum_{i=1}^{n} d_i = 2m$. To obtain a representation for this multigraph we specify a sequence of $2m$ stubs given as $(1^{d_1}, \ldots, n^{d_n})$ and let $\mathbf{X}$ be a random permutation of this stub sequence. The ordered pair (X_{2k-1}, X_{2k}) in this sequence is interpreted as an unordered site to which edge k is assigned. With the convention that (i, j) for $i \leq j$ is a canonical representation for the unordered pair, we define (Y_{2k-1}, Y_{2k}) as a canonical representation corresponding to (X_{2k-1}, X_{2k}), that is

$$(Y_{2k-1},Y_{2k}) = \begin{cases} (X_{2k-1},X_{2k}) & \text{if } X_{2k-1} \leq X_{2k} \\ (X_{2k},X_{2k-1}) & \text{if } X_{2k} \leq X_{2k-1} \end{cases} \tag{1}$$

for $k = 1,\ldots,m$. The sequence $\mathbf{Y} = (Y_1,\ldots,Y_{2m})$ is our canonical representation for the unordered multigraph obtained by random stub matching of a stub sequence that is distributed on the vertices according to a given degree sequence $\mathbf{d} = (d_1,\ldots,d_n)$. The distribution of its random edge multiplicities $\mathbf{M} = (M_{ij} : (i,j) \in R)$ is given in Theorem 1. These results are also given and discussed in Shafie (2012; 2015; 2016). An alternative to the edge multiplicity sequence $\mathbf{M}$ is the n by n edge multiplicity matrix with elements M_{ij}, where $M_{ji} = M_{ij}$ for all i and j in V. This matrix generalizes the adjacency matrix of simple graphs.

Theorem 1
Under RSM($\mathbf{d}$) the edge multiplicities have probabilities

$$P(\mathbf{M} = \mathbf{m}) = c2^{-t} , \tag{2}$$

where

$$c = \frac{m!2^m\mathbf{d}!}{(2m)!} = \frac{d_1!,\ldots,d_n!}{(2m-1)!!} , \tag{3}$$

and

$$t = m_1 + \sum\sum_{i\leq j} \log_2 m_{ij}! = \sum_{i=1}^{n} m_{ii} + \sum\sum_{i\leq j} \log_2 m_{ij}! . \tag{4}$$

Proof
The number of permutations of the stub sequence $\mathbf{X}$ is given by

$$\binom{2m}{\mathbf{d}} = \frac{(2m)!}{d_1!,\ldots,d_n!} . \tag{5}$$

In order to count how many of these that are interpreted as having $\mathbf{M} = \mathbf{m}$ we note that any of the m unordered pairs of stubs correspond to two ordered pairs if $X_{2k-1} \neq X_{2k}$ so that the numberof favourable cases is

$$\binom{m}{\mathbf{m}} 2^{m_2} , \tag{6}$$

where $m_2 = m - m_1$ and $m_1 = \sum_{i=1}^{n} m_{ii}$. Therefore

$$P(\mathbf{M} = \mathbf{m}) = \frac{\binom{m}{\mathbf{m}} 2^{m_2}}{\binom{2m}{\mathbf{d}}} = \frac{m!\, 2^m\, \mathbf{d}!}{(2m)!\, \mathbf{m}!\, 2^{m_1}} = c2^{-t} . \tag{7}$$

□

According to Theorem 1, the outcomes $\mathbf{m}$ with the same value $t = t(\mathbf{m})$ have the same probability, and these probabilities decrease with increasing values of t. In particular, simple graphs with no loops and no multiple edges are graphs with $t = 0$. If such graphs exist they have a common probability $m!2^m/\binom{2m}{\mathbf{d}}$. The statistic t can be used as a complexity measure that takes on positive values when there are loops or multiple edges at at least one site. Another such measure is $\sum_{i=1}^{n} m_{ii} + \sum\sum_{i\leq j} \binom{m_{ij}}{2}$, which is used by e.g. Janson (2009). Different complexity measures identify simple graphs by complexity 0 and other graphs by positive complexity. The complexity measure t has the property that it gives the same complexity to all graphs of the same probability under RSM. The random complexity measure

$$T = \sum_{i=1}^{n} M_{ii} + \sum\sum_{1\leq i\leq j\leq n} \log_2 M_{ij}! \quad . \tag{8}$$

is of special importance since it is related to the entropy of $\mathbf{M}$ as discussed in Section 4.

3 Local distributions of Loops and Edges

In this section we apply Theorem 1 to derive some new results on local distributions of loops and edges under RSM. In particular, we obtain local loop probabilities for M_{ii} in Theorem 2, and local probabilities for (M_{ii}, M_{ij}, M_{jj}) in Theorem 3. Local edge probabilities for M_{ij} are obtained as marginal sums in the trivariate distribution, and also, somewhat simpler, as marginal sums in the bivariate distribution of (M_{ii}, M_{ij}), which is given in Theorem 4.

Theorem 2
Under RSM$(\mathbf{d})$ the number of loops at vertex i has probability distribution

$$P(M_{ii} = u) = \frac{\binom{m}{u,\ d_i - 2u} 2^{d_i - 2u}}{\binom{2m}{d_i}} \tag{9}$$

for non-negative integers u satisfying

$$d_i - 2u \geq 0 \text{ and } m - d_i + u \geq 0 \ ,$$

that is

$$d_i - m \leq u \leq d_i/2 \ .$$

Proof
Consider the general probability that there are u loops at vertex i under RSM denoted by $P(M_{ii} = u)$ for $u = 0, \ldots, m$. To find how many of the $\binom{2m}{\mathbf{d}}$ possible stub sequences that generate u loops at i, arrange m edges with u loops at i, $d_i - 2u$ edges with the remaining i-stubs, and $m - d_i + u$ other edges. This number of arrangements is given by the multinomial coefficient $\binom{m}{u,\ d_i - 2u}$. The single i-stubs have two alternative locations in the $d_i - 2u$ edges. Finally, the remaining stubs are arranged in $\binom{2m - d_i}{\mathbf{d}^*}$ ways where $\mathbf{d}^*$ is the degree sequence $\mathbf{d}$ without d_i. This leads to

$$P(M_{ii} = u) = \frac{\binom{m}{u,\ d_i - 2u} 2^{d_i - 2u} \binom{2m - d_i}{\mathbf{d}^*}}{\binom{2m}{\mathbf{d}}} \ , \tag{10}$$

which simplifies to

$$P(M_{ii}=u)=\frac{m!\;d_i!\;(2m-d_i)!\;2^{d_i-2u}}{u!\;(d_i-2u)!\;(m-d_i+u)!\;(2m)!}\;. \tag{11}$$

□

Note that if $d_i \leq m$, then the possible values for M_{ii} are

$$u=0,1,\ldots,\lfloor d_i/2\rfloor\;.$$

If $d_i \geq m$, then the possible values are

$$u=d_i-m,d_i-m+1,\ldots,\lfloor d_i/2\rfloor\;,$$

so that there are only $\lfloor(2m-d_i)/2\rfloor+1$ instead of $\lfloor d_i/2\rfloor+1$ possible values of M_{ii}. The probability of no loops at vertex i, is given by Janson (2009) as

$$P(M_{ii}=0)=\prod_{j=1}^{d_i}\frac{2m-d_i-j+1}{2m-2j+1}\;. \tag{12}$$

The probability of no loops at vertex i using Equation (9) is equal to

$$P(M_{ii}=0)=\frac{\binom{m}{d_i}2^{d_i}}{\binom{2m}{d_i}}\;. \tag{13}$$

This formula can be developed according to the following which shows that it is equivalent to the expression of Janson (2009) for $P(M_{ii}=0)$ as a ratio between a falling factorial from $2m-d_i$ and a falling semi-factorial from $2m-1$, both carried out for d_i factors (in fact, d_i-1 factors suffice since the last one cancels):

$$\begin{aligned}P(M_{ii}=0)&=\frac{m!\;d_i!\;(2m-d_i)!\;2^{d_i}}{d_i!(m-d_i)!\;(2m)!}\\&=\frac{m!\;2^m\;(2m-d_i)!}{(2m)!\;(m-d_i)!\;2^{m-d_i}}\\&=\frac{(2m)!!\;(2m-d_i)!}{(2m)!\;(2m-2d_i)!!}\\&=\frac{(2m-2d_i-1)!!\;(2m-d_i)!}{(2m-1)!!\;(2m-2d_i)!}\\&=\frac{(2m-d_i)(2m-d_i-1)\cdots(2m-2d_i+1)}{(2m-1)(2m-3)\cdots(2m-2d_i+1)}\;.\end{aligned} \tag{14}$$

The distribution of M_{ii} can be summarized by its expected value

$$\mu_{ii}=E(M_{ii})=\sum_{k=1}^{m}P(Y_{2k-1}=Y_{2k}=i)=mQ_{ii} \tag{15}$$

and variance

$$\begin{aligned}\sigma_{ii}^2=Var(M_{ii})&=\sum_{k=1}^{m}\sum_{\ell=1}^{m}P(Y_{2k-1}=Y_{2k}=Y_{2\ell-1}=Y_{2\ell}=i)-\mu_{ii}^2\\&=\mu_{ii}(1-\mu_{ii})+m(m-1)Q_{iiii}\end{aligned}\;, \tag{16}$$

where

$$Q_{ii} = \frac{\binom{d_i}{2}}{\binom{2m}{2}} \text{ and } Q_{iiii} = \frac{\binom{d_i}{4}}{\binom{2m}{4}} \ . \tag{17}$$

We now consider dyad sites to get the trivariate distribution of $(M_{ii},\ M_{jj},\ M_{ij})$. The dyad site at vertices i and j has M_{ii} and M_{jj} loops and M_{ij} non-loops between i and j, $d_i - 2M_{ii} - M_{ij}$ external non-loops at i, $d_j - 2M_{jj} - M_{ij}$ external non-loops at j, and $m - d_i - d_j + M_{ii} + M_{jj} + M_{ij}$ remaining external edges. It can be considered as a multigraph on three vertices, the vertices i and j and a fictitious vertex aggregating the other vertices in V. The aggregated multigraph has degree sequence $(d_i,\ d_j,\ 2m - d_i - d_j)$ and edge multiplicity sequence

$$(M_{ii},\ M_{ij},\ d_i - 2M_{ii} - M_{ij},\ M_{jj},\ d_j - 2M_{jj} - M_{ij},\ m - d_i - d_j + M_{ii} + M_{jj} + M_{ij}) \ .$$

Theorem 3
Under RSM($\mathbf{d}$), the joint probability distribution of the edge multiplicities $(M_{ii},\ M_{jj},\ M_{ij})$ is given by

$$\begin{aligned} &P(M_{ii} = u,\ M_{jj} = v,\ M_{ij} = w) = \\ &= \frac{\binom{m}{u,\ v,\ w,\ d_i - 2u - w,\ d_j - 2v - w} 2^{d_i + d_j - 2u - 2v - w}}{\binom{2m}{d_i,\ d_j}} \end{aligned} \tag{18}$$

Proof
Applying the formula for the RSM model in Theorem 1 to the aggregated multigraph on three vertices gives

$$\begin{aligned} &P(M_{ii} = u,\ M_{jj} = v,\ M_{ij} = w) = \\ &= \frac{m!\ d_i!\ d_j!\ (2m - d_i - d_j)!\ 2^{d_i + d_j - 2u - 2v - w}}{u!\ v!\ w!\ (d_i - 2u - w)!\ (d_j - 2v - w)!\ (m - d_i - d_j + u + v + w)!\ (2m)!} \ , \end{aligned} \tag{19}$$

for possible outcomes given by non-negative integers u, v, w satisfying

$$d_i - 2u - w \geq 0,\ d_j - 2v - w \geq 0, \text{ and } m - d_i - d_j + u + v + w \geq 0 \ .$$

□

The distribution of M_{ij} is obtainable as a marginal distribution in the trivariate distribution given in Theorem 3 or, somewhat simpler, as a marginal distribution in the bivariate distribution given in the next theorem.

Theorem 4
Under RSM($\mathbf{d}$), the bivariate probability distribution of the edge multiplicities $(M_{ii},\ M_{ij})$ is given by

$$\begin{aligned} P(M_{ii}=u,\ M_{ij}=w) &= \frac{\binom{m}{u,\ w,\ d_i-2u-w}\binom{2m-2d_i+2u}{d_j-w}2^{d_i-2u}}{\binom{2m}{d_i,\ d_j}} \\ &= \frac{\binom{m}{u,\ m-d_i+u}\binom{d_i-2u}{w}\binom{2m-2d_i+2u}{d_j-w}2^{d_i-2u}}{\binom{2m}{d_i,\ d_j}}\ . \end{aligned} \tag{20}$$

Proof
If the probability in Equation (9) is written as a function of outcome and parameters according to $P(M_{ii}=u)=f(u|m,\ d_i)$, it follows that Equation (18) can be written as

$$\begin{aligned} &\frac{\binom{m}{u,\ w,\ d_i-2u-w}2^{d_i-2u}\binom{m-d_i+u}{v,\ d_j-2v-w}2^{d_j-2v-w}}{\binom{2m}{d_i,\ d_j}} \\ &= \frac{\binom{m}{u,\ w,\ d_i-2u-w}2^{d_i-2u}f(v|m-d_i+u,\ d_j-w)\binom{2m-2d_i+2u}{d_j-w}}{\binom{2m}{d_i,\ d_j}}\ . \end{aligned} \tag{21}$$

Now the arguments in f specifies a probability distribution for which Theorem 2 can be applied with modified parameters. Therefore, a summation over all v results in Equation (20). It might be noted that if the last expression in Equation (20) is summed over w we retain Equation (9). □

The distribution of M_{ij} can be summarized for $i<j$ by its expected value

$$\mu_{ij}=E(M_{ij})=mQ_{ij} \tag{22}$$

and variance

$$\sigma_{ij}^2=Var(M_{ij})=\mu_{ij}(1-\mu_{ij})+m(m-1)Q_{ijij}\ , \tag{23}$$

where

$$Q_{ij}=\frac{d_id_j}{\binom{2m}{2}} \quad \text{and} \quad Q_{ijij}=\frac{4\binom{d_i}{2}\binom{d_j}{2}}{\binom{2m}{2}\binom{2m-2}{2}}\ . \tag{24}$$

In order to judge multivariate spread of $(M_{ii},\ M_{jj},\ M_{ij})$ by relative entropy as defined in Section 4 we need to determine the total number of outcomes for the aggregated multigraph. This is given by a general formula for arbitrary RSM multigraphs on three vertices in the following section.

4 Aggregated Multigraphs on Triads

The random complexity measure T is of special importance since it is related to the entropy of $\mathbf{M}$ for the RSM($\mathbf{d}$) model. This is shown in Theorem 5 below, and it might be convenient to start by giving some definitions and preliminaries about entropies. The entropy $H(Z)$ for a discrete random variable Z with K outcomes z_k of positive probabilities $p(z_k)$ for $k = 1,\ldots,K$ is defined as

$$H(Z) = E\left(\log\frac{1}{p(Z)}\right) = \sum_{k=1}^{K} p(z_k)\log\frac{1}{p(z_k)} \ , \tag{25}$$

that is as the expected value of the so called uncertainty of Z, where uncertainty of an event is defined as the logarithm of its inverted probability. Uncertainty is usually measured in binary digits (bits) obtained with logarithms to the base 2 (so, for instance, an event of probability 1/8 is said to have uncertainty 3 bits). The entropy satisfies the inequalities

$$0 \leq H(Z) \leq \log_2 K \ ,$$

with equality to the left if Z has a single outcome and equality to the right if Z has a uniform probability distribution over its K outcomes. Entropy $H(Z)$ can therefore be used as a measure of spread between peakiness and flatness for the distribution of Z, or as a measure of uncertainty between certainty and maximal uncertainty for the outcomes of Z, such that $2^{H(Z)}$ approximates how many of the outcomes of Z are essential when uncertainty is taken into account. The proportion of essential bits $H(Z)/\log_2(K)$ measures relative uncertainty, and the number of non-essential bits $\log_2(K) - H(Z)$ measures discrepancy from uniformity. A more exact interpretation in terms of optimal binary prefix codes of the outcomes is provided by information theory. See, for instance, Cover & Thomas (2012). Multivariate entropies are useful in exploratory statistics and for statistical model testing as demonstrated in Frank (2000; 2011), and Frank & Shafie (2016).

Theorem 5
The entropy of the multiplicities $\mathbf{M}$ under RSM($\mathbf{d}$) depends on the random complexity measure T according to

$$H(M) = E(T) + \log_2(2m-1)!! - \sum_{i=1}^{n}\log_2(d_i!) \tag{26}$$

with expected complexity given by the local distributions of edge multiplicities according to

$$E(T) = \sum_{k=1}^{\lfloor d_1/2 \rfloor}(k + \log_2 k!)\sum_{i=1}^{n}P(M_{ii} = k) + \sum_{k=1}^{d_2}(\log_2 k!)\sum\sum_{1\leq i<j\leq n}P(M_{ij} = k) \tag{27}$$

where $d_1 \geq \cdots \geq d_n$.

Proof
According to Theorem 1 the expected uncertainty of $\mathbf{M}$ is equal to $H(\mathbf{M}) = E(T) - \log_2(c)$, where c is given by Equation (3). By using multiplicity counts

$$R_{1k} = \sum_{i=1}^{n} I(M_{ii} = k) \tag{28}$$

and

$$R_{2k} = \sum_{1 \le i < j \le n}\sum I(M_{ij} = k) \tag{29}$$

for $k = 0, 1, \ldots, m$, it follows that

$$T = \sum_{k=1}^{m} [(k + \log_2 k!)\ R_{1k} + \log_2 k!\ R_{2k}] \tag{30}$$

Since $M_{ii} \le \lfloor d_i/2 \rfloor$ and $M_{ij} \le \min(d_i, d_j)$, there are at most $\lfloor d_1/2 \rfloor$ positive terms in R_{1k} and at most d_2 in R_{2k} when degrees are ordered $d_1 \ge \cdots \ge d_n$. Hence the result for $E(T)$ follows. □

The calculations of entropy $H(\mathbf{M})$ and expected complexity $E(T)$ can be simplified if the needed distinct local distributions are obtained by using an algorithm that identifies all triads with distinct degree sequences $(d_i, d_j, 2m - d_i - d_j)$ and their distributions of M_{ii} and M_{ij}. Focus on aggregated triads can also be beneficial in order to develop sampling methods for multigraph inference. Assume, for instance, that data (m_{ii}, d_i) are available for vertices i in a vertex sample S_1 only, and not for all vertices in V. By Horvitz-Thompson estimators it is possible to estimate totals $E(R_{1k})$, m_1, and $2m$ for arbitrary sampling schemes with known inclusion probabilities for the vertices. A similar approach to a sample S_2 of vertex pairs with data $(m_{ii}, m_{ij}, m_{jj}, d_i, d_j)$ for $(i, j) \in S_2$, makes it possible to estimate totals $E(R_{1k})$, $E(R_{2k})$, m_1, m_2, and $2m$ that allow estimation of $E(T)$ and $H(\mathbf{M})$. In applications, such local sample data on loops, edges, and degrees might be observed and modelled with additive independent measurement errors, or might be partly known from other sources. Thus, sample data from aggregated triads suffice to estimate the global entropy and expected complexity of the multigraph.

The number $K(\mathbf{d})$ of outcomes of $\mathbf{M}$ is needed to judge spread and relative spread of M under RSM(d). It is no restriction to assume that $\mathbf{d}$ is ordered since $K(\mathbf{d})$ is invariant under permutations of the degrees in $\mathbf{d}$. Theorem 6 gives a general formula for arbitrary RSM multigraphs on three vertices with ordered degrees $a \ge b \ge c > 0$ and $a + b + c = 2m$. For $a \ge b > c = 0$ it was noted after Theorem 2 that multigraphs on two vertices have $K(a, b, 0) = K(a, b) = \lfloor b/2 \rfloor + 1$ outcomes.

Theorem 6

The number $K(a, b, c)$ of multigraphs on a triad with ordered degree sequence $a \ge b \ge c > 0$ is given by the following expressions when the integer parts of $a/2$, $b/2$, and $c/2$ are given by α, β, and γ:

$$\begin{aligned}
&K(2\alpha,2\beta,2\gamma) = (\beta+1)(\gamma+1)^2 - 2\binom{\gamma+2}{3} - \binom{\beta+\gamma-\alpha+2}{3}\\
&\text{ if } \alpha \geq \beta \geq \gamma \geq 1 \ ,\\
&K(2\alpha,2\beta+1,2\gamma+1) = (\beta+1)(\gamma+1)(\gamma+2) - 2\binom{\gamma+2}{3} - \binom{\beta+\gamma-\alpha+3}{3}\\
&\text{ if } \alpha > \beta \geq \gamma \geq 0 \ ,\\
&K(2\alpha+1,2\beta,2\gamma+1) = (\beta+1)(\gamma+1)(\gamma+2) - \binom{\gamma+2}{3} - \binom{\gamma+3}{3} - \binom{\beta+\gamma-\alpha+2}{3}\\
&\text{ if } \alpha \geq \beta > \gamma \geq 0 \ ,\\
&K(2\alpha+1,2\beta+1,2\gamma) = (\beta+1)(\gamma+1)^2 - \binom{\gamma+1}{3} - \binom{\gamma+2}{3} - \binom{\beta+\gamma-\alpha+2}{3}\\
&\text{ if } \alpha \geq \beta \geq \gamma \geq 1 \ .
\end{aligned} \tag{31}$$

Proof

For $a \geq b \geq c \geq 1$ and $a+b+c$ even there are four cases to consider:

$$\begin{aligned}
&1)\ a = 2\alpha,\ b = 2\beta,\ c = 2\gamma \ \text{ for } \ \alpha \geq \beta \geq \gamma \geq 1\\
&2)\ a = 2\alpha,\ b = 2\beta+1,\ c = 2\gamma+1 \ \text{ for } \ \alpha > \beta \geq \gamma \geq 0\\
&3)\ a = 2\alpha+1,\ b = 2\beta,\ c = 2\gamma+1 \ \text{ for } \ \alpha \geq \beta > \gamma \geq 0\\
&4)\ a = 2\alpha+1,\ b = 2\beta+1,\ c = 2\gamma \ \text{ for } \ \alpha \geq \beta \geq \gamma \geq 1 \ .
\end{aligned} \tag{32}$$

Let the edge multiplicities be denoted by x,y,z,u,v,w where x,y,z are the loop frequencies and u,v,w the non-loop frequencies satisfying

$$\begin{aligned}
2x+u+v &= a\\
2y+u+w &= b\\
2z+v+w &= c \ .
\end{aligned} \tag{33}$$

Due to these three restrictions we can express the six edge multiplicities by using only the loop counts and get the non-loop counts by

$$\begin{aligned}
u &= m-c+z-x-y\\
v &= m-b+y-x-z\\
w &= m-a+x-y-z \ .
\end{aligned} \tag{34}$$

Since all counts are non-negative this leads to the following inequalities valid for the possible integer values of x, y, z:

$$\begin{aligned}
&0 \le x \le a/2 \\
&0 \le y \le b/2 \\
&0 \le z \le c/2 \\
&x + y - z \le m - c \\
&x + z - y \le m - b \\
&y + z - x \le m - a \,.
\end{aligned} \tag{35}$$

If we rewrite this as restrictions on (x, y) for a fixed value of z according to

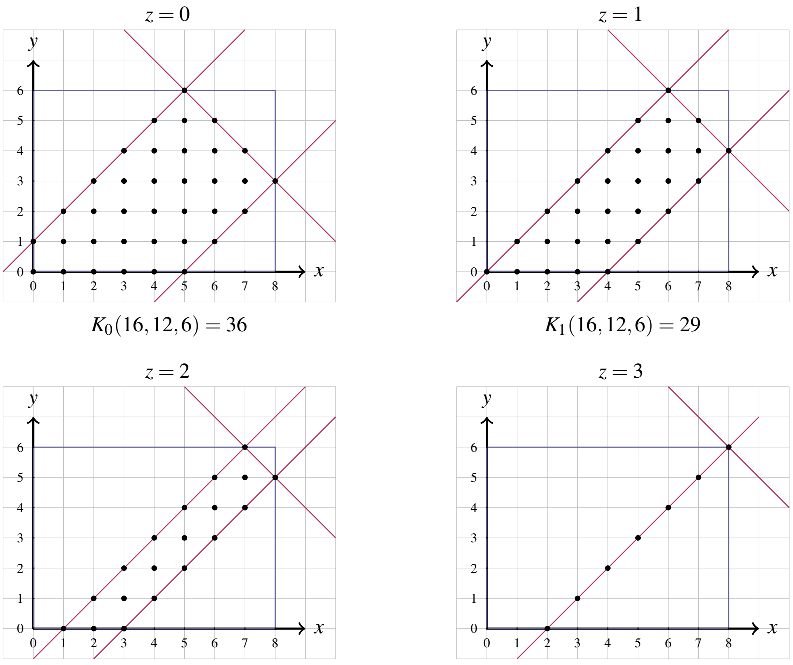
$$\begin{aligned}
&0 \le x \le a/2 \\
&0 \le y \le b/2 \\
&y \le C_z - x \ \text{ where } \ C_z = m - c + z \\
&B_z + x \le y \le A_z + x \ \text{ where } \ B_z = b - m + z \ \text{ and } \ A_z = m - a - z \,,
\end{aligned} \tag{36}$$

we can describe the possible (x, y) as the points with integer coordinates in the intersection of a rectangular region and the region below a line with negative slope and between two lines with common positive slope. If $K_z(a,b,c)$ denotes the number of possible points (x,y,z) for a fixed z, we get

$$K(a,b,c) = \sum_{z=0}^{\gamma} K_z(a,b,c) \tag{37}$$

as the total number of points in the $\gamma + 1$ slices of the three-dimensional region in (x,y,z)-space. Since the points counted in the (x,y)-plane belong to triangular and rectangular regions it is possible to use combinatorial formulas in Equation (31) to find $K_z(a,b,c)$ and sum these numbers to $K(a,b,c)$. □

Figure 1-4 show numerical examples to illustrate the four cases with different parity of the degrees.



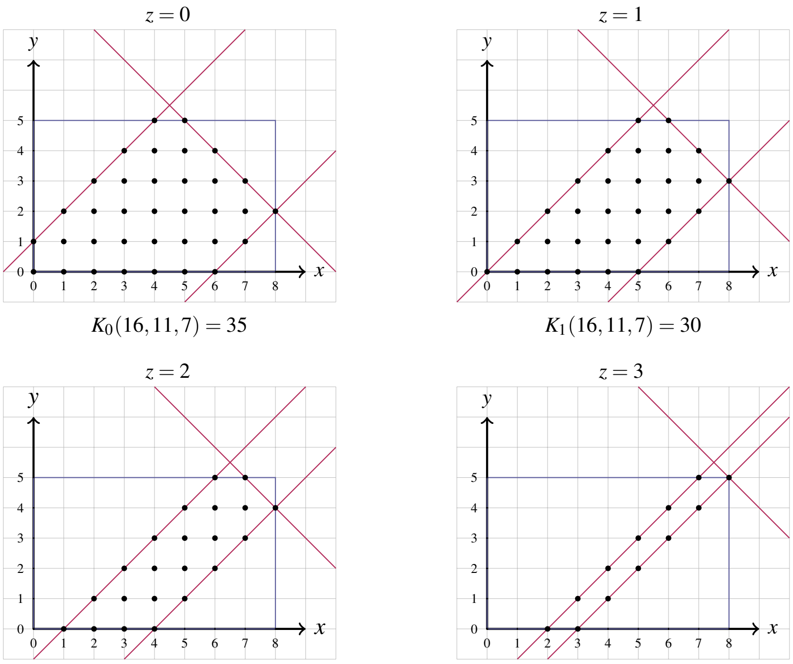
$z = 0$ $z = 1$

$K_0(16,12,6) = 36$ $K_1(16,12,6) = 29$

$z = 2$ $z = 3$

$K_2(16,12,6) = 19$ $K_3(16,12,6) = 7$

Fig. 1: Number of multiplicity sequences for degree sequence $(16,12,6)$ is $K(16,12,6) = 91$.



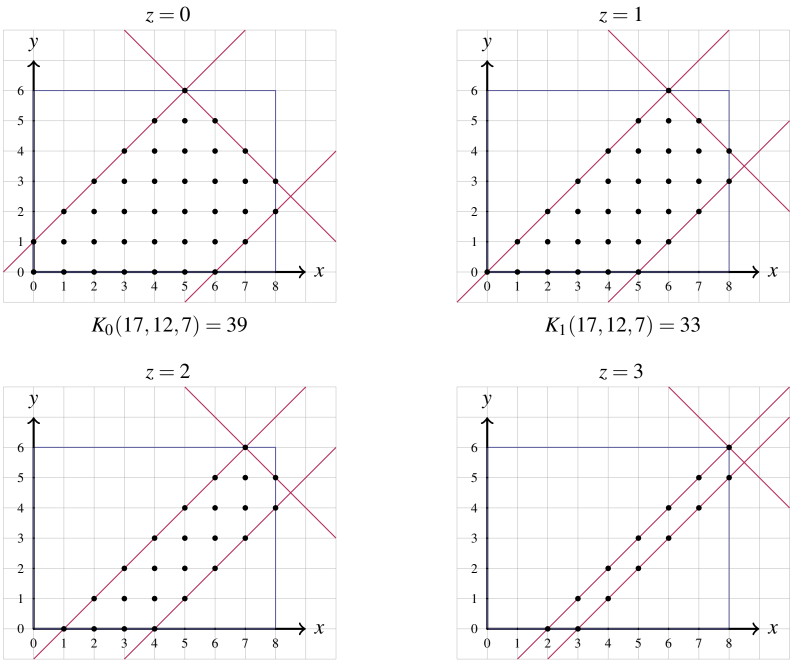
$z = 0$ $z = 1$

$K_0(16,11,7) = 35$ $K_1(16,11,7) = 30$

$z = 2$ $z = 3$

$K_2(16,11,7) = 22$ $K_3(16,11,7) = 12$

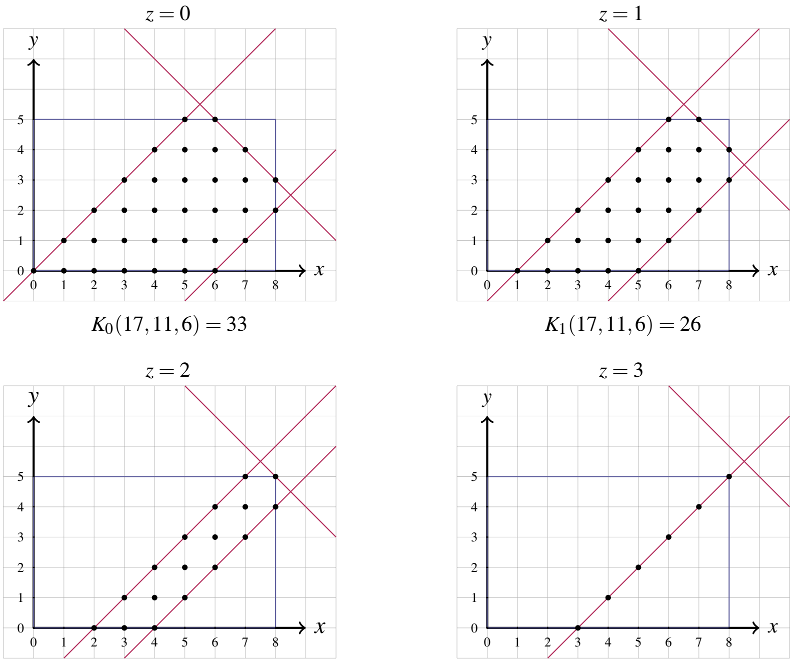
Fig. 2: Number of multiplicity sequences for degree sequence $(16,11,7)$ is $K(16,11,7) = 99$.



$K_0(17,12,7) = 39$ $K_1(17,12,7) = 33$

$K_2(17,12,7) = 24$ $K_3(17,12,7) = 13$

Fig. 3: Number of multiplicity sequences for degree sequence $(17,12,7)$ is $K(17,12,7) = 109$.



$K_0(17,11,6) = 33$ $K_1(17,11,6) = 26$

$K_2(17,11,6) = 17$ $K_3(17,11,6) = 6$

Fig. 4: Number of multiplicity sequences for degree sequence $(17,11,6)$ is $K(17,11,6) = 82$.

5 Some Illustrations

According to Theorem 5 the entropy of **M** is equal to

$$H(\mathbf{M}) = E(T) - \log_2(c) \ , \tag{38}$$

where T is the global complexity given by Equation (8) and the constant c is given by Equation (3). In this section we look at some numerical examples of how entropy $H(\mathbf{M})$, expected complexity $E(T)$, and $\log_2(c)$ vary for different n and $\mathbf{d}$ in regular and close to regular graphs. Consider first

$$\log_2(c) = m + \log_2(m!) - \log_2((2m)!) + \log_2(\mathbf{d}!) \ , \tag{39}$$

where $2m = d_1 + \cdots + d_n$. Stirling approximations

$$m! \approx \left(\frac{m}{e}\right)^m \sqrt{2\pi m} \tag{40}$$

can be used to all the factorials $m!$, $(2m)!$ and $\mathbf{d}! = (d_1! \cdot d_2! \cdots d_n!)$. The approximation to $\log_2(d!)$ is a function

$$f(d) = \log_2 \left[\left(\frac{d}{e}\right)^d \sqrt{2\pi d} \right] \ . \tag{41}$$

A Taylor expansion of this function at d_i around the average degree $\bar{d}$ gives

$$f(d_i) = f(\bar{d}) + (d_i - \bar{d})f'(\bar{d}) + \frac{(d_i - \bar{d})^2}{2} f''(\bar{d}) + \ldots \tag{42}$$

with an approximation given by its first three terms. By summation

$$\sum_{i=1}^{n} f(d_i) \approx n f(\bar{d}) + \sum_{i=1}^{n} \frac{(d_i - \bar{d})^2}{2} f''(\bar{d}) \ , \tag{43}$$

and an approximation of $\log_2(c)$ is thus given by

$$\log_2(c) \approx m + f(m) - f(2m) + n f(\bar{d}) + \frac{ns^2}{2} f''(\bar{d}) \ , \tag{44}$$

where s^2 is the degree variance and

$$f''(\bar{d}) = \frac{\log_2(e)}{\bar{d}} \left(1 - \frac{1}{2\bar{d}}\right) \ . \tag{45}$$

In order to consider how the expected complexity $E(T)$ depends on n and $\mathbf{d}$, we express Equation (8) as

$$E(T) = \sum_{i=1}^{n} E\left[M_{ii} + \log_2(M_{ii}!)\right] + \sum\sum_{1 \le i < j \le n} E\left[\log_2(M_{ij}!)\right] \tag{46}$$

and use Stirling approximations to the factorials as before. The approximation to $\log_2(M!)$ has a Taylor expansion

$$f(M) = f(\mu) + (M - \mu)f'(\mu) + \frac{(M - \mu)^2}{2} f''(\mu) + \ldots \tag{47}$$

so that an approximation is given by its three first terms and we get

$$E[f(M)] \approx f(\mu) + \frac{\sigma^2}{2} f''(\mu) \tag{48}$$

where $\mu = E(M)$ and $\sigma^2 = Var(M)$. Thus, we have an approximation of the expected complexity

$$\begin{aligned} E(T) &\approx \sum_{i=1}^{n} E[M_{ii} + f(M_{ii})] + \sum\sum_{1\leq i<j\leq n} E[f(M_{ij})] \\ &= \sum_{i=1}^{n} \left[\mu_{ii} + f(\mu_{ii}) + \frac{\sigma_{ii}^2}{2} f''(\mu_{ii})\right] + \sum\sum_{1\leq i<j\leq n} \left[f(\mu_{ij}) + \frac{\sigma_{ij}^2}{2} f''(\mu_{ij})\right] \end{aligned} \tag{49}$$

where means and variances are given by Equations (15) to (17) and Equations (22) to (24), and can be simplified for regular graphs according to

$$\mu_{ii} = E(M_{ii}) = \frac{\binom{d}{2}}{nd-1} \ , \tag{50}$$

$$\sigma_{ii}^2 = Var(M_{ii}) = \mu_{ii}(1-\mu_{ii}) + m(m-1)Q_{iiii} = \mu_{ii}(1-\mu_{ii}) + \frac{6\binom{d}{4}}{(nd-1)(nd-3)} \ , \tag{51}$$

$$\mu_{ij} = E(M_{ij}) = \frac{d^2}{nd-1} \ , \tag{52}$$

$$\sigma_{ij}^2 = Var(M_{ij}) = \mu_{ij}(1-\mu_{ij}) + m(m-1)Q_{ijij} = \mu_{ij}(1-\mu_{ij}) + \frac{d^2(d-1)^2}{(nd-1)(nd-3)} \ . \tag{53}$$

As a consequence,

$$E(T) \approx n\left[\frac{\binom{d}{2}}{nd-1} + g_1(n,d)\right] + \binom{n}{2} g_2(n,d) \tag{54}$$

where

$$g_1(n,d) = f(\mu_{ii}) + \frac{\sigma_{ii}^2}{2} f''(\mu_{ii}) \ , \tag{55}$$

$$g_2(n,d) = f(\mu_{ij}) + \frac{\sigma_{ij}^2}{2} f''(\mu_{ij}) \ . \tag{56}$$

The accuracy of the approximation of $E(T)$ can be obtained by comparing it to the exact value of $E(T)$ available from direct computation according to Equation (27) or according to

$$E(T) = \sum_u\sum_v\sum_w \left[n(u + \log_2(u!)) + \binom{n}{2}\log_2(w!)\right] P(M_{ii} = u, M_{jj} = v, M_{ij} = w) \ , \tag{57}$$

where M_{ii} and M_{ij} are loop and edge counts at and between any vertices i and j having degree d. The exact value of $E(T)$ requires the distribution of M_{ii} given in Theorem 2 and the marginal distribution of M_{ij} obtainable from the three-variate distribution given in Theorem 3 or from the bivariate distribution given in Theorem 4. The approximate value of $E(T)$ given in Equation (54) requires no more than the expected values and variances of the edge counts given by Equations (15), (16), (22) and (23).

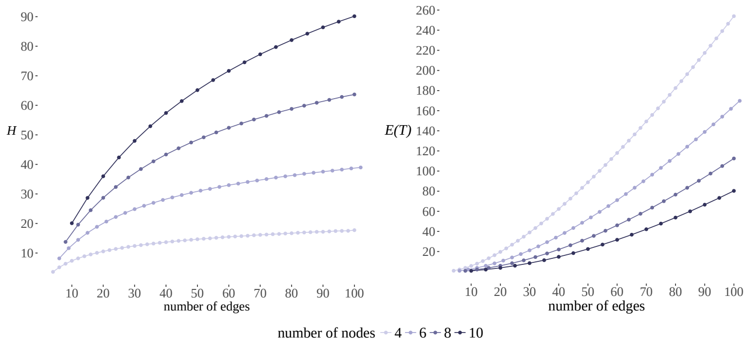


Fig. 5: Entropy H and expected complexity $E(T)$ for regular graphs with $n = 4, 6, 8, 10$, against $m = 4, 5, \ldots, 100$.

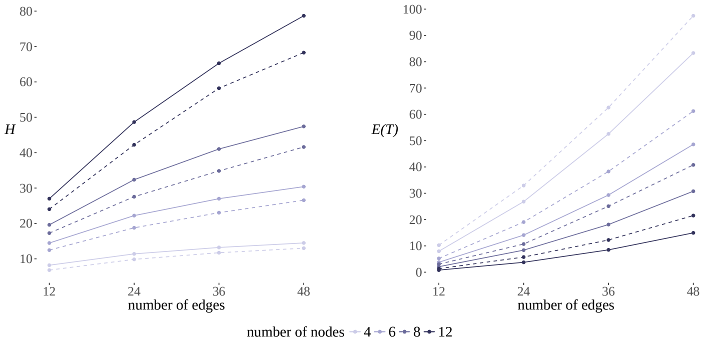


Fig. 6: Entropy H and expected complexity $E(T)$ for regular (solid) and irregular (dashed) graphs with $n = 4, 6, 8, 12$, against $m = 12, 24, 36, 48$.

Table 1 illustrates the accuracy of approximations of $E(T)$ and $H(\mathbf{M})$ based on second order Taylor expansion for graphs with $n = 8$ and $m = 24$, and degree variance $s^2 \leq 3$. As seen, both approximations are very good for all cases shown in Table 1.

Figure 5 and 6 illustrate how $E(T)$ and $H(\mathbf{M})$ vary for regular and irregular graphs with different combinations of n and m. For the irregular cases, the degree sequences are chosen to correspond to $2s = \bar{d}$, where $\bar{d} = 2m/n$, so that most of the degrees are within the interval from 0 to $2\bar{d}$.

We notice some interesting tendencies and stabilities in Figure 5 and 6. Both entropy and expected complexity increase with increasing number of edges. But with increasing number of vertices, entropy increases and expected complexity decreases. Thus, when further edges are distributed among a fixed number of sites of vertex pairs, entropy and expected complexity go up. When more sites are made available for a fixed number of edges, entropy goes up but expected complexity goes down. It is remarkable that these tendencies seem to remain for irregular as well as for regular degree sequences. These observations might inspire to investigations of various asymptotic tendencies in large multigraphs obtained by random stub matching.

Table 1: Exact and approximate values of $E(T)$ and $H(\mathbf{M})$ based on second order Taylor expansions for graphs with $n = 8$, $m = 24$, and with degree variance less than or equal to 3.

		Exact		Approximation	
Degree sequence	Degree variance	$E(T)$	$H(\mathbf{M})$	$\hat{E}(T)$	$\hat{H}(\mathbf{M})$
6 6 6 6 6 6 6 6	0	8.38	32.36	8.36	32.50
5 6 6 6 6 6 6 7	0.25	8.48	32.23	8.45	32.37
5 5 6 6 6 6 7 7	0.5	8.57	32.10	8.55	32.24
4 6 6 6 6 6 7 7	0.75	8.67	31.94	8.65	32.08
5 5 5 6 6 7 7 7	0.75	8.66	31.97	8.64	32.11
5 5 6 6 6 6 6 8	0.75	8.65	31.99	8.63	32.13
4 5 6 6 6 7 7 7	1	8.77	31.81	8.74	31.95
4 6 6 6 6 6 6 8	1	8.76	31.83	8.73	31.97
5 5 5 5 7 7 7 7	1	8.75	31.84	8.73	31.98
5 5 5 6 6 6 7 8	1	8.75	31.87	8.72	32.01
4 5 5 6 7 7 7 7	1.25	8.86	31.68	8.83	31.82
4 5 6 6 6 6 7 8	1.25	8.85	31.71	8.82	31.85
5 5 5 5 6 7 7 8	1.25	8.84	31.74	8.81	31.87
3 6 6 6 6 7 7 7	1.5	8.98	31.44	8.95	31.59
4 4 6 6 7 7 7 7	1.5	8.96	31.52	8.93	31.67
4 5 5 6 6 7 7 8	1.5	8.94	31.58	8.92	31.72
5 5 5 5 6 6 8 8	1.5	8.93	31.63	8.90	31.77
5 5 5 6 6 6 6 9	1.5	8.91	31.67	8.89	31.81
3 5 6 6 7 7 7 7	1.75	9.08	31.32	9.05	31.46
3 6 6 6 6 6 7 8	1.75	9.07	31.34	9.04	31.48
4 4 5 7 7 7 7 7	1.75	9.05	31.39	9.02	31.54
4 4 6 6 6 7 7 8	1.75	9.05	31.42	9.02	31.56
4 5 5 5 7 7 7 8	1.75	9.04	31.44	9.01	31.59
4 5 5 6 6 6 8 8	1.75	9.03	31.47	9.00	31.61
4 5 6 6 6 6 6 9	1.75	9.02	31.51	8.99	31.65
5 5 5 5 5 7 8 8	1.75	9.02	31.50	8.99	31.64
5 5 5 5 6 6 7 9	1.75	9.01	31.54	8.98	31.68
3 5 5 7 7 7 7 7	2	9.17	31.19	9.14	31.33
3 5 6 6 6 7 7 8	2	9.16	31.21	9.13	31.36
4 4 5 6 7 7 7 8	2	9.14	31.29	9.11	31.43
4 4 6 6 6 6 8 8	2	9.14	31.31	9.11	31.46
4 5 5 5 6 7 8 8	2	9.12	31.34	9.09	31.48
4 5 5 6 6 6 7 9	2	9.11	31.38	9.08	31.52
5 5 5 5 5 7 7 9	2	9.10	31.41	9.07	31.55
3 4 6 7 7 7 7 7	2.25	9.28	31.03	9.24	31.18
3 5 5 6 7 7 7 8	2.25	9.26	31.08	9.23	31.23
3 5 6 6 6 6 8 8	2.25	9.25	31.11	9.22	31.25
3 6 6 6 6 6 6 9	2.25	9.23	31.14	9.20	31.28
4 4 5 6 6 7 8 8	2.25	9.23	31.18	9.20	31.33
4 4 6 6 6 6 7 9	2.25	9.22	31.22	9.19	31.36
4 5 5 5 6 7 7 9	2.25	9.21	31.25	9.17	31.39
5 5 5 5 5 6 8 9	2.25	9.19	31.31	9.16	31.45
2 6 6 6 7 7 7 7	2.5	9.42	30.66	9.39	30.82
3 4 6 6 7 7 7 8	2.5	9.36	30.93	9.33	31.07
3 5 5 6 6 7 8 8	2.5	9.35	30.98	9.31	31.12
3 5 6 6 6 6 7 9	2.5	9.33	31.02	9.30	31.16
4 4 4 7 7 7 7 8	2.5	9.34	31.00	9.31	31.15
4 4 5 5 7 7 8 8	2.5	9.32	31.05	9.29	31.20
4 4 5 6 6 7 7 9	2.5	9.31	31.09	9.28	31.24
4 5 5 5 5 8 8 8	2.5	9.30	31.11	9.27	31.25
4 5 5 5 6 6 8 9	2.5	9.29	31.15	9.26	31.29
5 5 5 5 6 6 6 10	2.5	9.25	31.27	9.22	31.41
2 5 6 7 7 7 7 7	2.75	9.52	30.54	9.49	30.70
2 6 6 6 6 7 7 8	2.75	9.51	30.56	9.48	30.72
3 4 5 7 7 7 7 8	2.75	9.46	30.80	9.42	30.94
3 4 6 6 6 7 8 8	2.75	9.45	30.82	9.42	30.97
3 5 5 5 7 7 8 8	2.75	9.44	30.85	9.41	30.99
3 5 5 6 6 7 7 9	2.75	9.43	30.89	9.40	31.04
4 4 4 6 7 7 8 8	2.75	9.43	30.90	9.40	31.04
4 4 5 5 6 8 8 8	2.75	9.41	30.95	9.38	31.10
4 4 5 5 7 7 7 9	2.75	9.40	30.97	9.37	31.11
4 4 5 6 6 6 8 9	2.75	9.40	30.99	9.37	31.13
4 5 5 5 5 7 8 9	2.75	9.39	31.02	9.35	31.16
4 5 5 6 6 6 6 10	2.75	9.35	31.11	9.32	31.25
5 5 5 5 5 6 7 10	2.75	9.34	31.14	9.31	31.28
2 5 6 6 7 7 7 8	3	9.61	30.43	9.57	30.59
2 6 6 6 6 6 8 8	3	9.60	30.45	9.57	30.61
3 3 7 7 7 7 7 7	3	9.60	30.54	9.56	30.69
3 4 5 6 7 7 8 8	3	9.55	30.69	9.51	30.84
3 4 6 6 6 7 7 9	3	9.53	30.73	9.50	30.88
3 5 5 5 6 8 8 8	3	9.53	30.75	9.50	30.89
3 5 5 5 7 7 7 9	3	9.52	30.76	9.49	30.91
3 5 5 6 6 6 8 9	3	9.52	30.79	9.48	30.93
4 4 4 6 6 8 8 8	3	9.52	30.80	9.49	30.94
4 4 4 6 7 7 7 9	3	9.51	30.81	9.48	30.95
4 4 5 5 6 7 8 9	3	9.49	30.86	9.46	31.01
4 4 6 6 6 6 6 10	3	9.46	30.95	9.43	31.09
4 5 5 5 6 6 7 10	3	9.45	30.98	9.42	31.12
5 5 5 5 5 5 9 9	3	9.45	30.98	9.42	31.13

Acknowledgements

The research leading to these results has in part received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 319209.

References

Bender, Edward A, & Canfield, Rodney E. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory Series A*, **24**(3), 296–307.

Cover, Thomas M., & Thomas, Joy A. (2012). *Elements of information theory*. John Wiley & Sons.

Frank, Ove. (1988). Triad count statistics. *Annals of Discrete Mathematics*, **38**, 141–149.

Frank, Ove. (2000). Structural plots of multivariate binary data. *Journal of Social Structure*, **1**(4), 1–19.

Frank, Ove. (2011). Statistical information tools for multivariate discrete data. *Pages 177–190 of:* Pardo, Leandro, Balakrishnan, Narayanaswamy, & Gil, Maria Angeles (eds), *Modern mathematical tools and techniques in capturing complexity*. Springer Berlin Heidelberg.

Frank, Ove, & Shafie, Termeh. (2016). Multivariate entropy analysis of network data. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, **129**(1), 45–63.

Frank, Ove, & Strauss, David. (1986). Markov graphs. *Journal of the American Statistical Association*, **81**(395), 832–842.

Holland, Paul W., & Leinhardt, Samuel. (1976). Local structure in social networks. *Sociological Methodology*, **7**(1), 1–46.

Janson, Svante. (2009). The probability that a random multigraph is simple. *Combinatorics, Probability and Computing*, **18**(1–2), 205–225.

Kolaczyk, Eric D. (2009). *Statistical analysis of network data: Methods and models*. Springer Science & Business Media.

McKay, Brendan D., & Wormald, Nicholas C. (1991). Asymptotic enumeration by degree sequence of graphs with degrees $o(n^{1/2})$. *Combinatorica*, **11**(4), 369–382.

Shafie, Termeh. (2012). *Random multigraph – complexity measures, probability models and statistical inference*. Ph.D. thesis, Stockholm University.

Shafie, Termeh. (2015). A multigraph approach to social network analysis. *Journal of Social Structure*, **16**(1), 21.

Shafie, Termeh. (2016). Analyzing local and global properties of multigraphs. *Journal of Mathematical Sociology*, **40**(4), 239–264.

Wasserman, Stanley, & Faust, Katherine. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Wormald, Nicholas C. (1999). Models of random regular graphs. *London Mathematical Society Lecture Note Series*, 239–298.