Analyzing local and global properties of multigraphs

DOI:
10.1080/0022250X.2016.1219732

Document Version
Accepted author manuscript

Analysing Local and Global Properties of Multigraphs

Termeh Shafie

Department of Computer and Information Science
University of Konstanz, Germany
termeh.shafie@uni-konstanz.de

Abstract

The local structure of undirected multigraphs under two random multigraph models is analysed and compared. The first model generates multigraphs by randomly coupling pairs of stubs according to a fixed degree sequence so that edge assignments to vertex pair sites are dependent. The second model is a simplification that ignores the dependency between the edge assignments. It is investigated when this ignorance is justified so that the simplified model can be used as an approximation, thus facilitating the structural analysis of network data with multiple relations and loops. The comparison is based on the local properties of multigraphs given by marginal distribution of edge multiplicities and some local properties that are aggregations of global properties.

Keywords: configuration model, random stub matching, edge multiplicity, edge loop, complexity, multigraph aggregation.

1 Introduction

There is a long history of random graph models focusing on local and global statistics of network structure, and how well these statistics capture the main features and structural dependencies of actually observed networks. Some examples can be found in Erdős and Rényi (1959, 1960); Gilbert (1959); de Solla Price (1976); Frank and Strauss (1986); Watts and Strogatz (1998); Lusher et al. (2012), to name a few. For observed networks with binary symmetrical relations, these features and dependencies include a small diameter, high tendency for homophily, a high level of triadic closure, and a skewed degree distribution. More details on dependency forms and how well different graph models represent them can be found in Snijders (2011) and Robins (2013).

Less is known about the structural properties of network data consisting of more than just binary relations, and in particular, networks with edge loops. For these data structures, it is important to consider a multigraph representation and a modelling approach that allow for the possible occurrence of multiple edges and loops. Frank and Shafie (2016) introduce a novel way of using entropy tools to find inter-dependencies among a multi-dimensional set of network variables. This approach is particularly useful in multivariate networks consisting of a vertex set with at least one kind of edge, numerical or qualitative attributes defined on the vertices and edges, and in combination with other network statistics. Other existing methods and models for analysing multi-relational and multiple networks include multiplexity analysis (Lazega and Pattison, 1999; Koehly and Pattison, 2005), blockmodels (White et al., 1976; Fienberg et al., 1985; Nowicki and Snijders, 2001), algebraic models (Boorman and White, 1976; Pattison, 1993), network comparison using uniform random multiple graph distributions (Pattison et al., 2000), dyad independent models (Wasserman, 1987; Pattison and Wasserman, 2002) and exponential random (multi)graph models (Pattison and Wasserman, 1999;

Wang, 2012). Note however that by convention, loops are in many of the mentioned cases excluded from consideration. Following the approach presented in Shafie (2015), we consider multigraphs as undirected graphs where multiple edges and/or edge loops of different kinds are permitted and use a multigraph representation using a sequence of edge multiplicities at different vertex pair sites.

Multigraphs can appear naturally in various applications but can also be obtained using aggregation techniques applied to graphs. In this paper, these aggregations are exemplified and connected to the interplay of micro level properties of networks involving individual actors, and the macro level features where interest is in the occurrence of relationship and structural phenomena. Multigraphs are mainly concerned with macro levels, especially when obtained by aggregation based on vertex attributes. In other words, the aggregate macro level is of specific interest here, and using statistics under random multigraph models is a starting point for understanding and inferring structural features of aggregated empirical networks.

Network analysis has a tradition of investigating local dyads, triads, and higher order induced subgraphs to get insight into the network structure (Davis and Leinhardt, 1972; Holland and Leinhardt, 1970, 1976; Frank and Strauss, 1986; Frank, 1988). Local properties of a network that are generally valid for all local sites of a special kind can be called global properties. For instance, connectivity of all dyads and transitivity of all triads in a graph imply graph connectivity and graph transitivity. When a local property is not generally valid at different sites in a network, it might be of interest to use the property as a predictor or as a covariate of some other feature of interest in the network. For multigraphs, these local properties are e.g. number of loops at local vertex sites, number of multiple edges at local dyad sites, and more comprehensive local aggregations of global properties of the multigraph. Information on all these are contained in the edge multiplicity sequence which we specify according to a canonical order and investigate under two random multigraph models.

The first random multigraph model generates random multigraphs given a specified degree sequence that might be fixed or randomly generated according to independently and identically distributed degrees. This model assumes that the edges are formed by randomly coupling pairs of stubs (or semi-edges), so that the existence of an edge depends on the existence of others. These degree based processes are referred to as cumulative advantage processes by de Solla Price (1976), or more recently as preferential attachment by Barabási and Albert (1999). Although this model is referred to as the configuration or pairing model in the literature (Wormald, 1999; McKay and Wormald, 1990; Bender and Canfield, 1978), we introduce and refer to it as random stub matching (RSM). This is done for the following reasons. We are mainly interested in the generation of multigraphs, whereas most studies in the literature focus on the generation of simple graphs without multiple edges and loops. Moreover, most results on the mathematical properties of the configuration model are of the asymptotic kind (Janson, 2009; Bollobás, 1980; Wormald, 1980, 1981). The proportion of edges involved in either loops or multiple edges is vanishingly small as the number of vertices goes to infinity, and it is sometimes argued that we may discard or collapse them without much impact (Newman, 2003; Newman et al., 2003). We argue against this since there are applications where these properties are important to consider, for example when dealing with small multigraphs which can be obtained by the different mentioned aggregation procedures. Additionally, we introduce yet another aggregation technique in this article. This technique gives local aggregated multigraphs of order two and three which can be used to summarise some global information. Deriving exact formulas for different properties of multigraphs generated by stub matching is therefore of interest. Moreover, neglecting loops and multiple edges will not only discard valuable information that should be considered in many applications, it will also have a direct effect on the probabilities connected to the generation of graphs: each permutation of the stubs will no longer have equal probability of occurring and thus not all generated simple graphs appear with the same probability. Yet another reason for calling the model RSM is

that we use a new representation of multigraphs by its edge multiplicities which is obtained via an injective mapping from the permuted stub sequence to the edge multiplicity sequence.

The second multigraph model is introduced in Shafie (2015) and obtained by independent edge assignments (IEA) according to a common probability distribution over the vertex pair sites. This model assumes dyadic independence, i.e. no stochastic dependency exists between the multiple edges as they are assigned to sites and can be viewed as a generalisation of the Bernoulli random graph model for univariate networks. The independence assumption makes it more straightforward to derive and use statistics for analysing global network properties, as compared to the RSM model where similar derivations have higher combinatorial complexity due to the restrictions imposed by the degree sequence.

We analyse and compare the marginal distributions of edge multiplicities at local sites under IEA and RSM, and show that in order to facilitate the structural analysis, one model can be used to approximate the other given certain conditions. The comparison is made between the two models when the edge assignment probabilities in the IEA are chosen to correspond to those of the RSM defined by the fixed degrees. Thus, we are simply ignoring the stochastic dependence between edge assignment to obtain an approximate IEA model, which is preferred since it is more easily analysed than RSM. However, it is of interest to determine when this ignorance is justified. This is done by investigating the multiplicity distributions under IEA and RSM in two ways. First, a theoretical analysis of the central moments of the loop and non-loop multiplicity distributions is carried out and general results are derived. Second, numerical examples and simulations are used to indicate for which cases we have resemblance or discrepancy between the two distributions based on information divergence.

The outline of this article is as follows. In the next section, we present different possibilities of obtaining multigraph structures from network data, either as directly observed or as a product of different aggregation techniques. Further, the usefulness of random multigraph models for analysing such networks is discussed. In practice, it is often important to focus on local properties of the multigraphs. We describe how to investigate loop distributions at local vertices as well as distributions of loops and edges at local dyads of vertices. In Section 3 we introduce the two undirected random multigraph models IEA and RSM using a representation based on their edge multiplicities. Then under each model, the moments of the edge multiplicity distributions are given. In Section 4, we present methods of obtaining an approximate IEA model from an RSM model and focus on one such method in the remainder of the article. By comparing the central moments under the two models, general results are given for when RSM and approximate IEA have coinciding or different moments of local properties and how that is affected by the degrees. The comparison is also performed by simulations and numerical examples of how the divergence between the distributions of local edge frequency under RSM and approximate IEA varies for different degrees. In Section 5, the results from the marginal distributions of edge multiplicities are extended to analyse the global structure of multigraphs using the simultaneous distribution of edge multiplicities. The concept of simplicity and complexity of a multigraph is introduced and new statistics to analyse these properties are derived. In particular, it is shown that the multigraph distribution under RSM depends on a single complexity statistic. In the final section we summarise and conclude the results, and give some suggestions for related future topics.

2 Multigraph representation of network data

Network structures encountered in applications often consist of several different vertex attributes together with some properties or occurrences of events between pairs of vertices, which can be regarded as edge attributes. Multigraphs might appear more or less naturally

embedded in such structures, and it is important to be able to extract relevant multigraphs for instance in an exploratory analysis of the network.

In the following, we exemplify situations where multigraphs are obtained as directly observed or as a result from different aggregation techniques. These situations are discussed and linked to micro and macro level analysis of network data, i.e. how local compositional properties on vertex level lead to the emergence of global structural features.

Statistical models for network structure aim to formalise the interplay between local (micro) and structural global properties (macro), and to capture the interdependency among edges in the observed network. We motivate the use of random multigraph models and how they can be used to infer structural patterns in networks with multiple relations and loops. Notations are introduced and the local structure of multigraphs is conceptualised.

2.1 Motivation and examples

A multigraph that easily exhibits itself is given by several edges of different kinds mapped on the same vertex pair, e.g. different social interactions among a group of people, or an edge variable that counts occurrences of events at different sites of vertex pairs, e.g. number of calls within and between different departments of an organisation. These may be viewed as macro level data since observations are made on an aggregate level ignoring individual social actors. Questions in focus when analysing structural properties of such networks may be: does the presence of one social interaction among groups of people influence or explain the presence or absence of another form of interaction between them? Is there more within than between departmental communication in the organisation under study? The first question is concerned with the multiplicity of edges and the inter-dependencies among them. This is also called entrainment, multiplexity or interlocking (Koehly and Pattison, 2005; Pattison and Wasserman, 1999; White et al., 1976). The latter question is concerned with number of loops in the network and is referred to as homophily and the tendency for similar actors to connect to each other (McPherson et al., 2001; Snijders, 2011). These concerns are equally important to consider when micro level data is observed since they provide the key to finding local social processes that accumulate and form the network structure observed. This is illustrated below with a few examples.

Less natural multigraphs can be obtained by transformation of network data, e.g. social interactions when time is considered, appropriately chosen integer scales for edges in a valued graph, and aggregation based on vertex and/or edge attributes. Shafie (2015) provides a detailed description of all these different transformations with examples and a discussion on the possible occurrence of loops for the presented cases. We focus here on graphs transformed into multigraphs by partitioning vertices based on single or combined categories of vertex attributes, and with edges moving between and within these different categories. The aggregation is then from micro to macro level, where the aggregated final multigraph will have fewer vertices representing the attribute categories, and edge multiplicities representing relations between and within these categories. An exploratory or confirmatory analysis of the edge multiplicities will then indicate if social processes like homophily or entrainment are apparent. Statistics connected to these processes are introduced and discussed in Section 5.

With Figure 1 as reference, we give a few examples to illustrate aggregations based on vertex attributes. The original graph comprise of 25 vertices, 20 edges and four vertex attributes labelled A, B, C and D. This graph is aggregated into a final multigraph on four vertices representing the attribute categories. This example could for instance be communication within and between branches of a government (executive, legislative, judicial, and electoral), transmission of information within and between different company sectors (management, accounting, marketing, and commerce), or collegial co-operations within and between divisions of medicine (psychiatry, neurology, surgery and pharmacology). Note that this kind of ag-

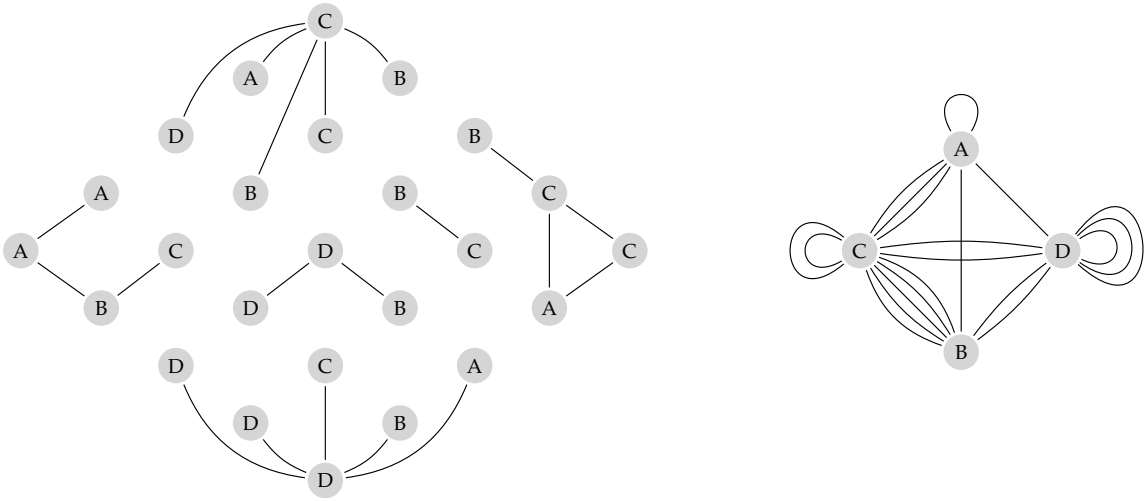

Figure 1: Initial graph (left) of 20 edges between 25 vertices with four attribute categories A, B, C and D. Final aggregated multigraph (right) on four new vertices corresponding to the attribute categories and with the same number of edges as the initial graph.

gregation is not restricted to only one edge type. The edges displayed in the initial graph of Figure 1 could be the aggregated result of different forms of communication, transmission or cooperation between the vertex categories, which with unaggregated edges may represent a directly observed multigraph.

The presented aggregation technique also has the capacity to capture the nested or hierarchical nature of most social science data. A simple example is given. Assume we are studying a network of friendships among school children categorised by gender (boy, girl) and living area (north, south). Combined categorisation of these attributes would thus result in the four vertices of the aggregated multigraph in Figure 1, and an analysis of this multigraph would reveal the affect gender and living area has on the formation of friendship ties at school.

2.2 Random multigraphs as models for network data

The presented simplification by aggregation is only useful if there are methods and models for understanding and inferring local and global structural properties of multigraphs. Probabilistic graph models allow for inference on structural properties by using the distribution of different network outcomes. Models found in the statistical literature aim to incorporate dependencies among edges in various ways; for example by different covariates, by conditioning on network statistics or assumed latent variables, and by representing dependencies among edges directly in the model. For a review of such models, see Snijders (2011) and Robins (2013). Another strain of models for network formation are strategic and use game theoretic tools. These approaches seek optimal network structures by maximising specified utility functions on social actors or agents (see e.g. Jackson (2005)). Common to these modelling schemes is the specification of micro level processes that generate macro level structures via the different dependence assumption on how edges between pairs of vertices occur (i.e. transitivity, homophily, multiplexity, etc.). These processes are formulated so that they generate structures similar to those observed in social networks. However, when it comes to assessing the fit of models to network data, the literature is mainly focused on micro level processes in simple or multi-relational graphs, without any loops. As seen from previous section, multigraph are mainly concerned with the macro level, either by direct observation or by data aggregation. Therefore, a new branch of models need to be considered in this context, or revisited to check

compatibility to other networks than simple ones (as done in Section 3). In other words, it is of interest to study how network statistics under multigraph models capture phenomena observed in empirical macro level networks with multiple edges and loops.

We focus here on uniform distributions for families of multigraphs which we refer to as random multigraphs. Comparing empirical network features to those expected under the specified random multigraph models is a good starting point for inferring whether observed outcomes are consistent with the estimated expected ones, while also checking the reliability of the estimates via their variance. Thus, we can detect deviation from randomness which informs on processes that are dominant in the formation of the network, and we can determine how these processes characterise the relations between vertices. This kind of baseline comparison will then provide guidance in future model specifications. We start by introducing some notations used throughout this paper.

2.3 Notations

A finite graph g with n labelled vertices and m labelled edges associates with each edge an ordered or unordered vertex pairs. Let $V = \{1, \ldots, n\}$ and $E = \{1, \ldots, m\}$ be the set of vertices and edges labelled by integers, and let R denote the set of available sites for the edges. For directed graphs the site space is $R = V \times V$ and the number of sites is given by $r = n^2$. Let the sequence $\mathbf{X} = (X_1, \ldots, X_{2m})$ with $\mathbf{X} \in R^m$ represent a general directed multigraph with m edges, where (X_{2k-1}, X_{2k}) is the site of edge k for $k = 1, \ldots, m$. For undirected graphs, we consider (i, j) with $i \leq j$ as a canonical representation for the unordered vertex pairs. Thus, the site space is $R = \{(i, j) \in V \times V : i \leq j\}$ and the number of sites is $r = \binom{n+1}{2}$. Let the sequence $\mathbf{Y} = (Y_1, \ldots, Y_{2m})$ represent an undirected multigraph where $(Y_{2k-1}, Y_{2k}) \in R$ is defined as (X_{2k-1}, X_{2k}) if $X_{2k-1} \leq X_{2k}$, and as (X_{2k}, X_{2k-1}) otherwise, for $k = 1, \ldots, m$. The sequence $\mathbf{Z} = (Z_1, \ldots, Z_{2m})$ is obtained from $\mathbf{Y}$ by ordering the edges non-decreasingly so that the edge sites are listed in the canonical order

$$(1,1) < (1,2) < \cdots < (1,n) < (2,2) < (2,3) < \cdots (2,n) < \cdots < (n,n) \ ,$$

with

$$(Z_1, Z_2) \leq (Z_3, Z_4) \leq \cdots \leq (Z_{2m-1}, Z_{2m}) \ .$$

Thus, the edge sequence $\mathbf{Z}$ represents the vertex labelled graph given by $\mathbf{Y}$, but without the edge labels.

A multigraph with labelled vertices and undistinguished edges is represented by its edge multiplicity sequence $\mathbf{M} = (M_{ij} : (i, j) \in R)$ where the edge multiplicity M_{ij} denotes the number of edges at different sites $(i, j) \in R$. There is a one-to-one correspondence between sequences $\mathbf{Z}$ and $\mathbf{M}$. The edge multiplicity counts are given by

$$M_{ij} = \sum_{k=1}^{m} I(Z_{2k-1} = i, Z_{2k} = j) = \sum_{k=1}^{m} I(Y_{2k-1} = i, Y_{2k} = j) \ .$$

The number of loops at vertex i is denoted M_{ii} and the number of edges between vertices i and j is denoted M_{ij} where $i < j$. The edge multiplicity sequence $\mathbf{M}$ has total

$$\sum_{i \leq j} \sum M_{ij} = m \ ,$$

and the degree of vertex i, which also can be interpreted as the number of edge stubs or half edges at vertex i, is given by

$$d_i = \sum_{j=i}^{n} M_{ij} + \sum_{j=1}^{i} M_{ji} \ ,$$

where $i = 1, \ldots, n$. It is also convenient to consider the edge multiplicities as entries in a matrix

$$\mathbf{M} = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ 0 & M_{22} & \cdots & M_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_{nn} \end{bmatrix}$$

and obtain the degrees as the sums of the entries in a row and the entries in the same column in $\mathbf{M}$, or equivalently as row sums (or column sums) in the matrix sum of $\mathbf{M}$ and its transpose $\mathbf{M}'$ which is given by

$$\mathbf{M} + \mathbf{M}' = \begin{bmatrix} 2M_{11} & M_{12} & \cdots & M_{1n} \\ M_{12} & 2M_{22} & \cdots & M_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ M_{1n} & M_{2n} & \cdots & 2M_{nn} \end{bmatrix} .$$

This matrix can be viewed as the equivalence of the adjacency matrix of simple graphs. The degree sequence $\mathbf{d} = (d_1, \ldots, d_n)$, which is also referred to as the stub multiplicity sequence, has total $\sum_{i=1}^{n} d_i = 2m$.

2.4 Local structure of Multigraphs

We look at the local structure of multigraphs from two perspectives. One is where local edge sites are equipped with count statistics that inform on global properties when summed over all vertices in the multigraph. The simplest local site of a multigraph is a vertex with its number of loops. In an undirected multigraph on vertex set $V = \{1, \ldots, n\}$, this is denoted M_{ii} for vertex $i \in V$. The local structure of a dyad site comprising the two distinct vertices i and j in V consists of their numbers of loops, M_{ii} and M_{jj}, together with the number of edges between them, M_{ij}. Local properties of the n vertex sites and the $\binom{n}{2}$ dyad sites are network properties that can be used as exploratory tools in network analysis. Their distribution provide information about how the local properties vary over the network and can be of interest as explanatory factors of other network variables or for comparison between networks.

Another approach for analysing the local structure of multigraphs is to view the local property as summaries of global properties in a multigraph. Here we consider yet another aggregation procedure where a fictitious vertex aggregates all the other vertices not part of the local site under study. In the following, we describe local aggregated multigraphs of order two and three.

The local structure at vertex i is specified by the edge multiplicity sequence $(M_{ii}, d_i - 2M_{ii}, m - d_i + M_{ii})$ for the m edges distributed as loops, non-loops, and external edges at vertex i. This sequence can be considered as a multiplicity sequence of an aggregated multigraph on two vertices with degree sequence $(d_i, 2m - d_i)$: vertex i and a fictitious vertex denoted $\bar{i}$, aggregating all other vertices in V. In other words, $\bar{i}$ is the complement of vertex i in V. Then, the non-loops at i are edges between vertex i and $\bar{i}$, and the external edges at vertex i are loops at vertex $\bar{i}$.

Similarly, the dyad site at vertices i and j with M_{ii} and M_{jj} loops and M_{ij} non-loops at 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. The aggregated multigraph has degree sequence $(d_i, d_j, 2m - d_i - d_j)$ and edge multiplicity sequence $(M_{ii},\ M_{ij},\ d_i - 2M_{ii} - M_{ij},\ M_{jj},\ d_j - 2M_{jj} - M_{ij},\ m - d_i - d_j + M_{ii} + M_{jj} + M_{ij})$. It can be considered as a multigraph on three vertices: i, j and a fictitious vertex denoted $\overline{ij}$ aggregating the other vertices in V. Then, external non-loops at i and j are edges between vertex i and $\overline{ij}$, and between j and $\overline{ij}$, respectively, and the remaining external edges are loops at vertex $\overline{ij}$. These

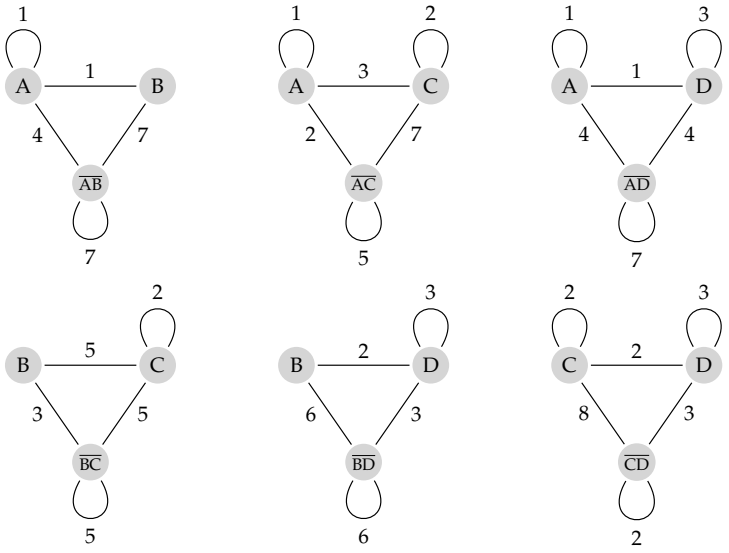

Figure 2: Aggregated multigraphs obtained from dyads selected from the multigraph in Figure 1. The vertex $\overline{ij}$ sums over all other vertices except the selected i and j in the multigraph, and edge labels indicate the edge multiplicities.

kinds of local aggregated multigraphs obtained from dyads selected in Figure 1 are shown in Figure 2, where the fictitious vertex $\overline{ij}$ is the complement of vertices i and j in V, and edge labels indicate the edge multiplicities. Sample data of this kind can be used to estimate global properties and latent (unobserved) local properties of a network, but this is out of the scope of this article. We focus here on the marginal distribution of M_{ij} for $i \leq j$ which can be summarised by its expected value and variance. In the next section, we introduce the two random multigraph models under which we study and compare these distributions under.

3 Random Multigraph Models

A random multigraph is given by a probability distribution over some class of multigraphs. Following Shafie (2015), we represent a multigraph by its sequence of edge multiplicities and use the edge multiplicity distribution to present two special kinds of random multigraph models. The first model is the configuration model (Bender and Canfield, 1978) revisited as a random multigraph model following a uniform multigraph distribution given a degree sequence. The second model can be viewed as a Bernoulli random graph model generalised by Shafie (2015) to apply for multigraphs.

3.1 Random Stub Matching

The first model generates random undirected multigraphs given a fixed degree sequence $\mathbf{d} = (d_1, \ldots, d_n)$ where $\sum_{i=1}^{n} d_i = 2m$. In order to get a multigraph with a fixed degree sequence we interpret it as a multiplicity sequence of $2m$ stubs or half edges. The m edges are then obtained by randomly matching the stubs of which d_i are attached to vertex $i \in V$. Since the assignment of edges is given by matching the stubs of the vertices, the model is called random stub matching model and denoted RSM($\mathbf{d}$), or shortly just RSM.

The representation of an RSM multigraph is obtained as follows. A sequence of $2m$ stubs is specified such that the first d_1 are equal to 1, the next d_2 are equal to 2, etc. A convenient notation is $(1^{d_1}, 2^{d_2}, \ldots, n^{d_n})$. Let $\mathbf{X}$ be random permutation of $(1^{d_1}, 2^{d_2}, \ldots, n^{d_n})$, which earlier was defined as a representation for a general directed multigraph. The ordered

pair (X_{2k-1}, X_{2k}) in this sequence is interpreted as an unordered site to which edge k is assigned. In order to obtain our canonical edge sequence $\mathbf{Y}$, we modify $\mathbf{X}$ by vertex shifts according to earlier description and obtain unordered vertex pairs (i, j) where $i \leq j$. This gives a general representation of an undirected multigraph with labelled edges where we note that the unordered pairs (Y_{2k-1}, Y_{2k}) are identically but not independently distributed. The multiplicity sequence $\mathbf{M} = (M_{ij} : (i, j) \in R)$ is then a representation of multigraph with undistinguishable edges for $\mathbf{Y} \sim \mathrm{RSM}(\mathbf{d})$, and the distribution of the edge multiplicities M_{ij} can be summarised by its first two central moments given in Theorem 1.

Theorem 1. *Let M_{ij} denote the random edge multiplicity at vertex pair (i, j) in the canonical site space $R = \{(i, j) : 1 \leq i \leq j \leq n\}$. The expected value is given by*

$$E_{\mathrm{RSM}}(M_{ij}) = \begin{cases} \dfrac{\binom{d_i}{2}}{(2m-1)} & \text{for } i = j \\ \dfrac{d_i d_j}{(2m-1)} & \text{for } i < j \ , \end{cases}$$

and the variance by

$$V_{\mathrm{RSM}}(M_{ij}) = \begin{cases} \dfrac{\binom{d_i}{2}}{2m-1}\left(1 - \dfrac{\binom{d_i}{2}}{2m-1}\right) + \dfrac{6\binom{d_i}{4}}{(2m-1)(2m-3)} & \text{for } i = j \\ \dfrac{d_i d_j}{(2m-1)}\left(1 - \dfrac{d_i d_j}{2m-1}\right) + \dfrac{d_i d_j (d_i - 1)(d_j - 1)}{(2m-1)(2m-3)} & \text{for } i < j \ . \end{cases}$$

Proof. Let $S(\mathbf{d})$ be the set of sequences that are permutations of the sequence of $2m$ stubs, obtained from the degree sequence $\mathbf{d}$. The number of possible permutations is given by

$$|S(\mathbf{d})| = \binom{2m}{\mathbf{d}} = \frac{(2m)!}{\mathbf{d}!} = \frac{(2m)!}{d_1! \cdots d_n!} \ ,$$

and each randomly permuted stub sequence $\mathbf{X}$ appears with equal probability

$$\frac{1}{\binom{2m}{\mathbf{d}}} \ .$$

Let $\mathbf{d}^*$ denote the degree sequence $\mathbf{d}$ after fixing two stubs from i and j in $\mathbf{X}$. The number of permutations of $S(\mathbf{d}^*)$ is thus given by

$$|S(\mathbf{d}^*)| = \binom{2m-2}{\mathbf{d}^*} \ .$$

The probability of coupling stubs to edges in $\mathbf{X}$ can then be written as

$$\frac{\binom{2m-2}{\mathbf{d}^*}}{\binom{2m}{\mathbf{d}}} \ , \tag{1}$$

where the nominator depends on whether the removal of two stubs are both from vertex i, or one stub is from i and one is from j. These two cases can be rewritten from Equation (1) as

$$P_{ij} = P\left((X_{2k-1}, X_{2k}) = (i, j)\right) = \begin{cases} \dfrac{\binom{d_i}{2}}{\binom{2m}{2}} & \text{for } i = j \\ \dfrac{d_i d_j}{2m(2m-1)} & \text{for } i \neq j \ , \end{cases}$$

where $\sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij} = 1$. Following earlier definition, the sequence $\mathbf{Y}$ is obtained from $\mathbf{X}$ by vertex shifts according to $Y_{2k-1,2k} = (\min(X_{2k-1}, X_{2k}), \max(X_{2k-1}, X_{2k}))$ for $k = 1, \ldots, m$, so that the number of $\mathbf{X}$ yielding the same $\mathbf{Y}$ is only affected for $i \neq j$. Thus, the probability of undirected edges in our canonical representation for the unordered multigraph $\mathbf{Y}$ corresponding to $\mathbf{X}$ is equal to

$$Q_{ij} = P\left((Y_{2k-1}, Y_{2k}) = (i,j)\right) = \begin{cases} P_{ii} = \frac{\binom{d_i}{2}}{\binom{2m}{2}} & \text{for } i = j \\ 2P_{ij} = \frac{d_i d_j}{\binom{2m}{2}} & \text{for } i < j \\ 0 & \text{for } i > j \ . \end{cases} \tag{2}$$

The edge multiplicities can be represented as

$$M_{ij} = \sum_{k=1}^{m} I_{ijk} \ .$$

where

$$I_{ijk} = I\left((Y_{2k-1}, Y_{2k}) = (i,j)\right) = \begin{cases} 1 & \text{if } (Y_{2k-1}, Y_{2k}) = (i,j) \\ 0 & \text{otherwise} \ , \end{cases} \tag{3}$$

for $(i,j) \in R$ and $k = 1, 2, \ldots, m$. The expected value of the indicator variables is $E_{\text{RSM}}(I_{ijk}) = Q_{ij}$ so that

$$E_{\text{RSM}}(M_{ij}) = mQ_{ij} \ ,$$

and substituting Q_{ij} from Equation (2) for $i = j$ and $i < j$ gives the first moment expression.

In order to find the variance of M_{ij}, we need the covariance between I_{ijk} and $I_{ij\ell}$ which can be given as

$$Cov_{\text{RSM}}(I_{ijk}, I_{ij\ell}) = \begin{cases} Q_{ij}(1 - Q_{ij}) & \text{for } k = \ell \\ Q_{ijij} - Q_{ij}^2 & \text{for } k \neq \ell \ , \end{cases}$$

where Q_{ij} is given in Equation (2) and Q_{ijij} are the probabilities of coupling four stubs into two edges in the sequence $\mathbf{Y}$. With a similar reasoning as before, we let $\mathbf{d}^*$ denote the degree sequence $\mathbf{d}$ after fixing four stubs in $\mathbf{X}$. The number of permutations of $S(\mathbf{d}^*)$ is now given by

$$|S(\mathbf{d}^*)| = \binom{2m-4}{\mathbf{d}^*} \ ,$$

and depends on whether the removal is of four i stubs, or two i and two j stubs. Thus, the probabilities $P(Y_{2k-1} = Y_{2k} = Y_{2\ell-1} = Y_{2\ell} = i)$ for $i = j$, and $P(Y_{2k-1} = Y_{2\ell-1} = i, Y_{2k} = Y_{2\ell} = j)$ for $i < j$ are given by

$$Q_{ijij} = \begin{cases} \frac{\binom{d_i}{4}}{\binom{2m}{4}} & \text{for } i = j \\ \frac{4\binom{d_i}{2}\binom{d_j}{2}}{\binom{2m}{2}\binom{2m-2}{2}} & \text{for } i < j \ . \end{cases} \tag{4}$$

From this it follows that

$$\begin{aligned} V_{\text{RSM}}(M_{ij}) &= \sum_{k=1}^{m}\sum_{\ell=1}^{m} Cov_{\text{RSM}}(I_{ijk}, I_{ij\ell}) \\ &= mQ_{ij}(1 - Q_{ij}) + m(m-1)(Q_{ijij} - Q_{ij}^2) \\ &= mQ_{ij}(1 - mQ_{ij}) + m(m-1)Q_{ijij} \ , \end{aligned}$$

and the second moment expression is obtained by substituting Q_{ij} from Equation (2) and Q_{ijij} from Equation (4), for $i = j$ and $i < j$, respectively. □

3.2 Independent Edge Assignments

The second model is basic random multigraph model which can be viewed as a simple site selection model. This model is introduced in Shafie (2015) and obtained by independent edge assignment to vertex pairs (i, j) in the canonical site space $R = \{(i, j) : 1 \leq i \leq j \leq n\}$. Assume that the edge assignment probabilities are denoted by $\mathbf{Q} = (Q_{ij} : (i, j) \in R)$ satisfying $Q_{ij} \geq 0$ and $\sum\sum_{(i,j)\in R} Q_{ij} = 1$, and assign m labelled edges $k = 1, \ldots, m$ independently to vertex pairs $(Y_{2k-1}, Y_{2k}) \in R$. Thus, the multigraph with labelled edges is represented by the sequence $\mathbf{Y} = (Y_1, \ldots, Y_{2m})$, and the multigraph with undistinguishable edges is represented by the sequence of edge multiplicities $\mathbf{M} = (M_{ij} : (i, j) \in R)$. The assumptions imply that $\mathbf{M}$ is multinomial distributed with parameters m and $\mathbf{Q}$ for $\mathbf{Y} \sim \text{IEA}(m, \mathbf{Q})$, where $\text{IEA}(m, \mathbf{Q})$ is the notation used for this independent edge assignment model. Thus, the marginal distribution of M_{ij} is binomially distributed with parameters m and Q_{ij}, and we directly get that

$$E_{\text{IEA}}(M_{ij}) = mQ_{ij}$$

and

$$V_{\text{IEA}}(M_{ij}) = mQ_{ij}(1 - Q_{ij}) \ .$$

4 Approximations and Comparisons

Shafie (2015) describes how some special statistics of simplicity and complexity can be used to analyse the global structure of multigraphs under the $\text{IEA}(m, \mathbf{Q})$ model. These statistics inform on the structural signature of a network but are more complicated to derive under models that consider interdependencies among the occurrence of edges, e.g. when conditioning on the degrees. Thus, it is of interest to approximate the distribution of $\mathbf{M}$ with an $\text{IEA}(m, \mathbf{Q})$ model and to know when this approximation is justified. We will return to the global structure and the mentioned special statistics in Section 5. In this section, we present two ways of obtaining an approximate $\text{IEA}(m, \mathbf{Q})$ model from an $\text{RSM}(\mathbf{d})$ model, and focus on the second approximation method for further investigation in the remainder of this article. This investigation is performed by comparing the distributions of edge multiplicities under the two models to find out when the IEA approximation is appropriate to apply. In other words, the local moment properties of the edge multiplicities presented in this section are used to compare the two models and determine when one model can be used to approximate the other, thus making it more feasible to analyse the global features of an observed network.

4.1 Approximations

A Bayesian version of the $\text{RSM}(\mathbf{d})$ model is obtained by assigning a prior to the parameter $\mathbf{d}$, i.e. assuming that the stubs are independently attached to the n vertices according to a probability distribution $\mathbf{p} = (p_1, \ldots, p_n)$ with $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$. This implies that $\mathbf{d}$ is the outcome of a random degree sequence $\mathbf{D}$ that is multinomial distributed with parameters $2m$ and $\mathbf{p}$. The multinomial distribution can thus be viewed as a Bayesian model for the stub frequencies. It also follows that the multiplicity sequence $\mathbf{M}$ has an $\text{IEA}(m, \mathbf{Q})$ distribution with edge probability sequence $\mathbf{Q}(\mathbf{p})$. With a slight abuse of notation, not used elsewhere in this article, we denote this probability sequence $\mathbf{Q} = (Q_{ij} : (i, j) \in R)$ where

$$Q_{ij} = \begin{cases} p_i^2 & \text{for } i = j \\ 2p_i p_j & \text{for } i < j \ . \end{cases}$$

Another way to obtain an IEA$(m, \mathbf{Q})$ model is to simply ignore the dependency between the edge assignments in the RSM$(\mathbf{d})$ model. This implies that the distribution of $\mathbf{M}$ is approximated with the edge probability sequence $\mathbf{Q}$ defined as a function of the fixed degrees $\mathbf{d}$ under RSM$(\mathbf{d})$ according to Theorem 1. These assignment probabilities are given in equation Equation (2). This model is denoted IEA$(m, \mathbf{Q}(\mathbf{d}))$ and can be viewed as repeated assignments with replacements of stubs, whereas RSM$(\mathbf{d})$ is repeated assignments without replacement of stubs. We will in the following focus on this second method of approximation and determine for which cases the dependency of edge assignments are negligible. This is done by a comparison of the moments of the edge multiplicities theoretically, and by simulations of the edge multiplicity distributions under the two models. Note that when we in the following use the shorthand notation IEA, we refer to the approximate IEA$(m, \mathbf{Q}(\mathbf{d}))$ model.

4.2 Comparison of the moments of edge multiplicities

In this section we present some general results for the centrality and spread of the edge multiplicity distributions under RSM$(\mathbf{d})$ and IEA$(m, \mathbf{Q}(\mathbf{d}))$ using their first and second moments given in Section 3.

The comparison of the variance can conveniently use that

$$V_{\text{RSM}}(M_{ij}) = V_{\text{IEA}}(M_{ij}) + \Delta_{ij} \quad \text{for } i \leq j \ ,$$

where the difference between the variances is

$$\Delta_{ij} = m(m-1)(Q_{ijij} - Q_{ij}^2) \ , \tag{5}$$

and Q_{ijij} and Q_{ij} are given in Equation (4) and Equation (2), respectively. This expression is used to show for which values of d_i and d_j the variance of the IEA distribution is smaller or larger than the variance of the RSM distribution. For $i = j$, we determine the sign of Δ_{ii} for values of d_i where $2 \leq d_i \leq 2m-1$. For $i < j$, we set $a = \min(d_i, d_j)$ and $b = \max(d_i, d_j)$ and determine the sign of Δ_{ij} for different pairs of values (a, b) with $1 \leq a \leq b$ and $a + b \leq 2m$ for $m > 1$.

Theorem 2 states that the edge multiplicity distributions under RSM and IEA have common expected values and negative variance differences, except for some degenerate and special cases.

Theorem 2. *Let M_{ij} denote the random edge multiplicity at vertex pair (i, j) in the canonical site space $R = \{(i, j) : 1 \leq i \leq j \leq n\}$. The following statements hold:*

(i) $E_{\text{RSM}}(M_{ij}) = E_{\text{IEA}}(M_{ij})$.

(ii) $V_{\text{RSM}}(M_{ii}) < V_{\text{IEA}}(M_{ii})$, *unless the degenerate cases* $d_i = 1$ *and* $d_i = 2m$ *are valid with* $V_{\text{RSM}}(M_{ii}) = V_{\text{IEA}}(M_{ii}) = 0$.

(iii) $V_{\text{RSM}}(M_{ij}) < V_{\text{IEA}}(M_{ij})$, *unless the special case is valid where* d_i *and* d_j *lie symmetrically around* m *and are given by* $m \pm k$ *for some non-negative integer*

$$k < \sqrt{\frac{m(m-1)}{4m-3}} \ .$$

This special case has $V_{\text{RSM}}(M_{ij}) > V_{\text{IEA}}(M_{ij})$.

Proof. (i) Follows directly from the proof of Theorem 1 and the fact that $\mathbf{Q}$ is a function of $\mathbf{d}$ in the approximate IEA model.
(ii) By inserting Equation (2) and Equation (4) into Equation (5) for $i = j$, we get

$$\Delta_{ii} = m(m-1)Q_{ii}\left[\frac{(d_i-2)(d_i-3)}{(2m-2)(2m-3)} - \frac{d_i(d_i-1)}{2m(2m-1)}\right] \ .$$

To see that the second term in the bracket is larger than the first term, we notice that

$$\frac{d_i - k}{2m - k} = 1 - \frac{2m - d_i}{2m - k}$$

is a decreasing function of k. From this it follows that $\Delta_{ii} < 0$ for $1 < d_i < 2m$. Further, for $d_i = 1$ and $d_i = 2m$ we have that $Q_{ii} = 0$ and $Q_{ii} = 1$, respectively, both leading to $\Delta_{ii} = 0$ and $V_{\text{IEA}}(M_{ii}) = 0$.
(iii) Let $a = \min(d_i, d_j)$ and $b = \max(d_i, d_j)$. By inserting Equation (2) and Equation (4) into Equation (5) for $i < j$, we get

$$\Delta_{ij} = m(m-1)Q_{ij}\left[\frac{(a-1)(b-1)}{\binom{2m-2}{2}} - \frac{ab}{\binom{2m}{2}}\right] ,$$

which has the same sign as the function

$$f(a,b) = \frac{(a-1)(b-1)}{ab} - \frac{\binom{2m-2}{2}}{\binom{2m}{2}} = \left(1 - \frac{1}{a}\right)\left(1 - \frac{1}{b}\right) - (1-\theta) ,$$

where $\theta = (4m-3)/m(2m-1)$ and $0 < \theta < 1$. Now $f(a,b) < 0$ for $1 \le a \le b \le m-1$, and $f(1,b) < 0$ for $1 \le b \le 2m-1$. For fixed value a or fixed value b, $f(a,b)$ is increasing in the other variable. Moreover, $f(m,m) > 0$. In order to find the critical curve between positive and negative values of $f(a,b)$, we set $f(a,b) = 0$ and solve for b to get

$$b = \frac{a-1}{a\theta - 1} .$$

The intersection between this curve and the upper boundary $b = 2m - a$ of the (a,b)-region defined by $1 \le a \le b$ and $a + b \le 2m$ is obtained as the solution to the quadratic equation

$$a^2 - 2ma + \frac{2m-1}{\theta} = 0$$

with roots

$$a = m \pm \sqrt{\frac{m(m-1)}{4m-3}} .$$

The relevant root is $m - \sqrt{m(m-1)/(4m-3)}$ since $a = \min(d_i, d_j)$ cannot be larger than m. It follows that

$$f(a, 2m-a) < 0 \quad \text{for} \quad 1 \le a < m - \sqrt{\frac{m(m-1)}{4m-3}} ,$$

$$f(a, 2m-a) > 0 \quad \text{for} \quad m - \sqrt{\frac{m(m-1)}{4m-3}} < a \le m ,$$

and

$$f(a, 2m-a) = 0 \quad \text{if} \quad a = m - \sqrt{\frac{m(m-1)}{4m-3}} .$$

With a similar investigation of the line $b = 2m - 1 - a$ and the critical curve, we find no intersection and therefore $f(a,b) < 0$ for $1 \le a \le b \le 2m - 1 - a$. It follows that

$$\Delta_{ij} > 0 \quad \text{only for} \quad m - \sqrt{\frac{m(m-1)}{4m-3}} < a = 2m - b \le m ,$$

that is for the $\left\lceil \sqrt{m(m-1)/(4m-3)} \right\rceil$ integer points $(a, 2m-a)$ with

$$m - \sqrt{\frac{m(m-1)}{4m-3}} < a \leq m$$

on the upper boundary. Moreover, $\Delta_{ij} < 0$ for the other $m^2 - \left\lceil \sqrt{m(m-1)/(4m-3)} \right\rceil$ points (a, b) in the (a, b)-region. □

In Figure 3, a numerical illustration of the proof of Theorem 2 is shown.

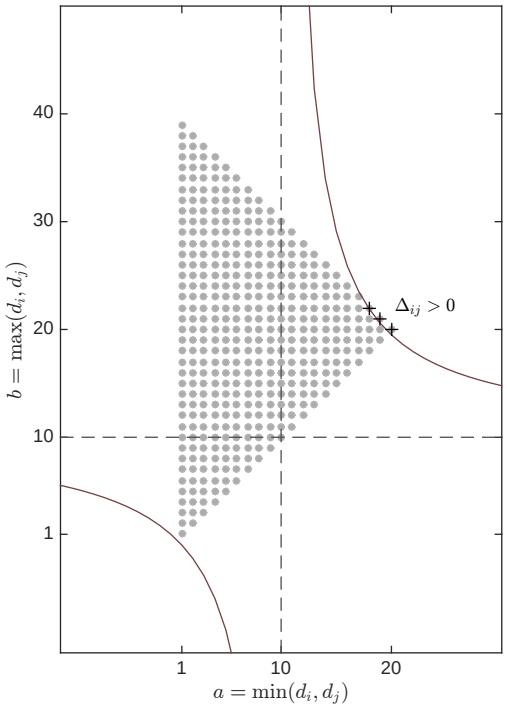

Figure 3: The points represent possible degree pairs (a, b) at a given vertex pair in a graph with $m = 20$ edges. A critical curve is separating points with positive and negative variance difference Δ_{ij} between the edge multiplicity distributions obtained at (a, b) by random stub matching and by independent edge assignments. The points with positive Δ_{ij} are marked with $(+)$.

Corollary 1. *For $1 \leq i < j \leq n$, $V_{\mathrm{RSM}}(M_{ij})$ is maximal for $k = 0$ and decreases for increasing k.*

Corollary 2. *For $1 \leq i < j \leq n$, it is only for the special case $n = 2$ that $V_{\mathrm{RSM}}(M_{ij}) > V_{\mathrm{IEA}}(M_{ij})$.*

As stated in Corollary 2, it is only when we have two vertices that the IEA approximation gives smaller variance than that of an RSM model for edge multiplicities M_{ij} where $i < j$. This occurs when a and b lie at the same distance from m, and this distance is strictly less than $\sqrt{m(m-1)/(4m-3)}$. Thus, $\Delta_{ij} > 0$ for only one choice $(a, 2m-a) = (m, m)$ if $m < 5$, two choices (m, m) and $(m-1, m+1)$ if $5 \leq m < 17$, three choices if $17 \leq m < 37$, four choices if $37 \leq m < 65$, five choices if $65 \leq m < 101$, and so forth. Of the m cases of $(a, 2m-a)$, only $\left\lceil \sqrt{m(m-1)/(4m-3)} \right\rceil$ have a variance larger than $V_{\mathrm{IEA}}(M_{ij})$, so even if the number of cases increases with increasing m, the proportion of cases decreases towards zero. This is illustrated in Figure 4, where we also notice that the proportion is not monotonically decreasing.

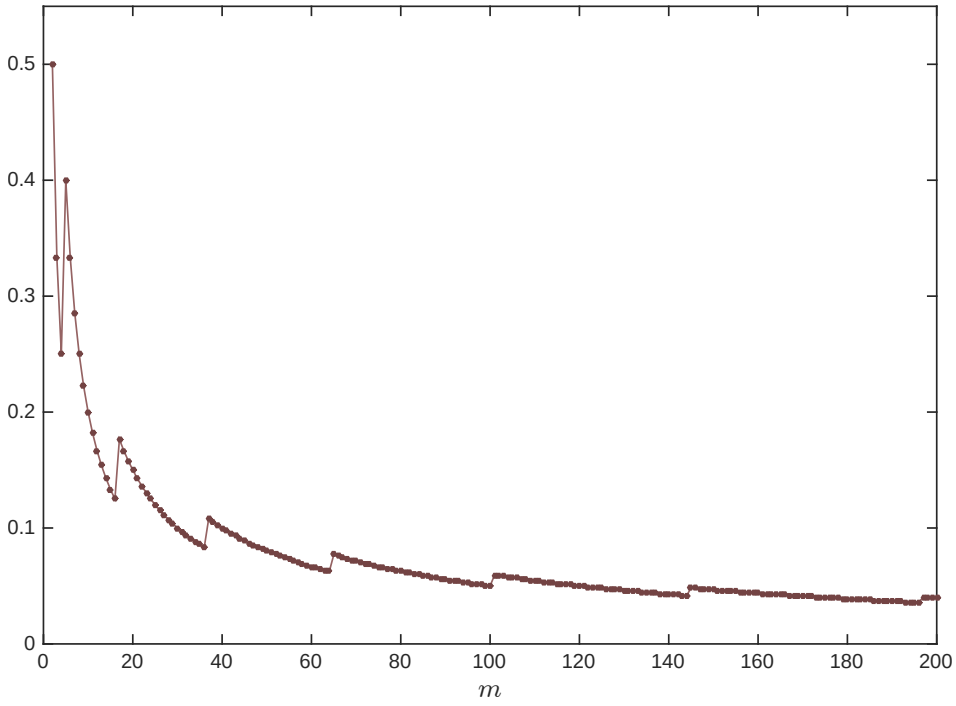

Figure 4: Proportion of the degree pairs $(a, 2m-a)$ for $a = 1, \ldots, m$ with edge multiplicity variance larger for random stub matching than for independent edge assignments. The proportions are plotted against edge frequency m.

4.3 Comparison by simulation of edge multiplicity distributions

Using simulations of the edge multiplicity distributions under RSM($\mathbf{d}$) and IEA($m, \mathbf{Q}(\mathbf{d})$), we can study and compare them to each other. This allows us to study the differences between them, thus giving us an idea of when an IEA approximation is reasonable to adopt. The comparison is done using information divergence which is a measure of discrepancy between the two distributions. Let $\mathbf{P} = (P_0, \ldots, P_m)$ and $\mathbf{B} = (B_0, \ldots, B_m)$ denote the RSM and IEA distribution of M_{ij}, respectively. Note that we keep this notation although we do distinguish the cases $i = j$ and $i < j$ in the following. The divergence can be interpreted as the number of additional binary digits required when an optimal binary code is used for the outcomes obtained from $\mathbf{B}$ rather than $\mathbf{P}$. The divergence between $\mathbf{P}$ and $\mathbf{B}$ is formally given by

$$D(\mathbf{P}, \mathbf{B}) = \sum_{\substack{k=0 \\ P_k>0}}^{m} P_k \left[\log \frac{1}{B_k} - \log \frac{1}{P_k}\right] = \sum_{\substack{k=0 \\ P_k>0}}^{m} P_k \log \frac{P_k}{B_k} \,, \tag{6}$$

which is an expected log-likelihood ratio. The log-likelihood ratios can be of any sign but their weighted sum, the divergence $D(\mathbf{P}, \mathbf{B})$, is non-negative and zero only when there is no discrepancy between the two distributions. For more details on these and similar information theoretic tools, see Frank and Shafie (2016); Frank (2011); Gray (2011); Kullback (1968).

We start with an illustration of how the divergence between the probability distributions of edge frequency M_{ij} for $i < j$ under RSM and under IEA vary for different numbers of stubs at vertex i and vertex j, i.e. for different ordered degree pairs (a, b) with $1 \leq a \leq b$ and $a + b \leq 2m$ where m is the total edge frequency. The case $m = 20$ is illustrated in Figure 5 where divergence $D(\mathbf{P}, \mathbf{B})$ is plotted against degree pairs (a, b) using a colour coding of standardised divergence values applied to the unit squares located around points

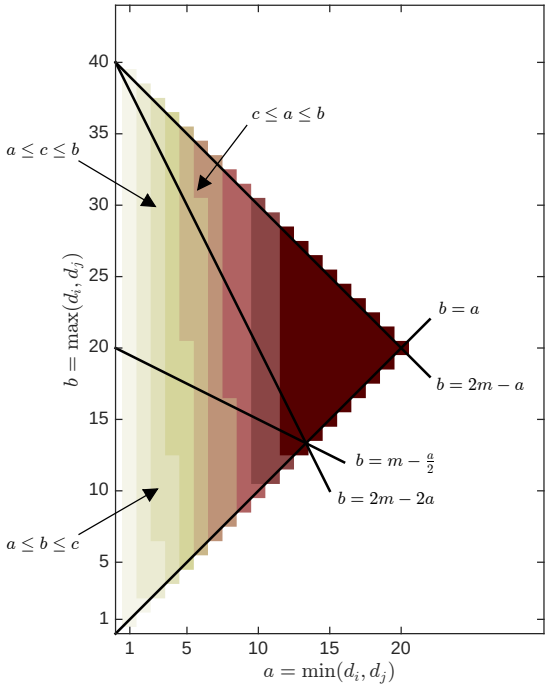

Figure 5: Divergence between the multiplicity distribution under random stub matching and under independent edge assignments for edges between two vertices with ordered degree pair (a, b) when total edge frequency is $m = 20$ and total number of stubs is $2m = a + b + c$. A darker colour at the unit squares located around points (a, b) represent a larger divergence than a brighter colour.

(a, b). The divergences for all possible degree pairs (a, b) are calculated and their maxima are determined. Standardised divergence values are obtained by dividing with the maxima. Every 10th percentile of this standardised distribution is then calculated and assigned a colour where darker colours represent higher divergences, i.e. darker colours are assigned to unit squares where the RSM distribution deviates the most from the IEA distribution. Letting $c = 2m - a - b$ denote the number of stubs at other vertices than i and j, border lines are drawn in Figure 5 where c is equal to the stub frequencies a and b. These two border lines $b = 2m - 2a$ and $b = m - a/2$, together with the border lines $b = 2m - a$ and $b = a$, divide the figure in three regions corresponding to whether c is smaller than, or larger than, or between the two stub frequencies a and b. The upper region in Figure 5 represents cases where $c \leq a \leq b$. Here, we have the majority of the high divergence values implying that the RSM distribution and the IEA distribution deviates the most. The middle region in Figure 5 represents cases where $a \leq c \leq b$. Here, the majority of the region has a brighter colour implying less deviation between the RSM distribution and the IEA distribution. The same applies for the lower region in Figure 5 which represents cases where $a \leq b \leq c$. Here, even less deviation is seen between the two distributions. Thus we can conclude that the more stubs we have at other vertices than i and j, the more resemblance we have between the distributions of M_{ij} under RSM and IEA.

In Figure 6, similar illustrations of divergence are shown together with figures showing the variance difference Δ_{ij} between the edge multiplicity distributions obtained at ordered degree pairs (a, b). This allows us to see for which values of Δ_{ij} the divergence is close to zero, thus indicating more resemblance between the two distributions. The divergences shown in the left column of Figure 6 follow the same colour coding as described above for Figure 5. The variance differences shown in the right column of Figure 6 are colour coded according

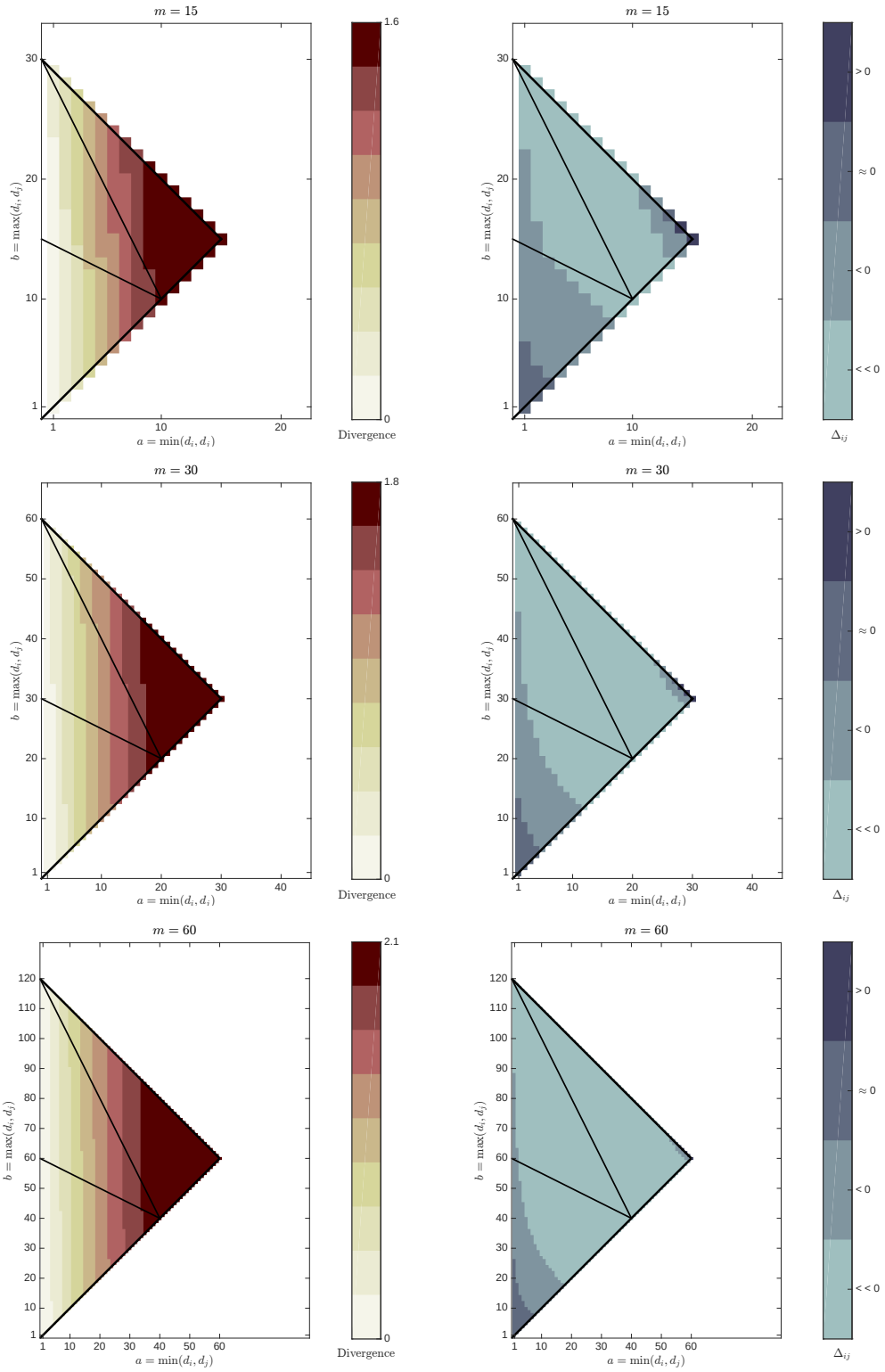

Figure 6: Divergence and variance difference Δ_{ij} between the multiplicity distribution under random stub matching and under independent edge assignments for edges between two vertices with ordered degree pair (a, b) when total edge frequency is m=15, 30 and 60, and total number of stubs is $2m$. The colour at the unit squares located around points (a, b) represent divergence (left) and variance difference (right) colour coded according to shown legends for each case. The three regions divided by border lines correspond to those in Figure 5.

to the following. The variance differences at degree pairs (a, b) are calculated, rounded to the first decimal and assigned to one of four different labelled categories: if $\Delta_{ij} < -0.5$ then"$<< 0$", if $-0.5 \leq \Delta_{ij} < 0$ then "< 0", if $\Delta_{ij} = 0.0$ then "≈ 0", and if $\Delta_{ij} > 0.0$ then "> 0". Note that the negative variance difference is split into two categories since Theorem 2 showed that the majority of the Δ_{ij} fall below zero. From Figure 6 we note that the divergence between the distributions under IEA and RSM is low for $\Delta_{ij} \approx 0$, i.e. when $V_{\text{RSM}}(M_{ij}) \approx V_{\text{IEA}}(M_{ij})$. This occurs when almost all stubs are at other vertices than i and j. Moreover, the largest discrepancy, as measured by the divergence, is seen for $\Delta_{ij} > 0$, i.e. when $V_{\text{RSM}}(M_{ij}) > V_{\text{IEA}}(M_{ij})$. We also note from Figure 6 that as m increases, the maximal divergence value increases but the region (i.e. the number of ordered degree pairs (a, b)) with low divergence and approximate zero variance difference also increases. Therefore, it is hard to note any clear effects of m increasing. However, the region for low divergence and a small negative variance difference does match each other better as m increases. We can conclude the following: the distribution of edge frequency M_{ij} for $i < j$ under RSM and under IEA resemble each other the most when there is a large number of stubs at other vertices than the vertex pair under study and when $V_{\text{RSM}}(M_{ij}) \approx V_{\text{IEA}}(M_{ij})$, i.e. when $\Delta_{ij} \approx 0$. Further, as m increases, more resemblance between the two distributions are apparent when $V_{\text{RSM}}(M_{ij}) < V_{\text{IEA}}(M_{ij})$, i.e. when $-0.5 < \Delta_{ij} < 0$.

We perform a similar investigation of the simulated loop distributions M_{ii}, that is for the number of loops at vertex i, under RSM and under IEA. Note that the range of the loop multiplicity distribution under IEA is $k = 0, \ldots, m$, while the range under RSM is smaller. Under RSM, $d_i \leq m$ and the possible values for M_{ii} are $k = 0, \ldots, \lfloor d_i/2 \rfloor$ so that there are only $\lfloor d_i/2 \rfloor + 1$ possible values as compared to $m + 1$ under IEA. Thus, the RSM range proportion is $(\lfloor d_i/2 \rfloor + 1)/(m+1)$ of the IEA distribution, and the divergence in Equation (6) is calculated for this proportion only (the weighted sum of the log-likelihood ratios is over $P_k > 0$ only). Figure 7 shows how the divergence, given as coloured bars, varies for different stub proportion $d_i/2m$ (or range proportion $(\lfloor d_i/2 \rfloor + 1)/(m + 1)$) for $m = 15$, 30 and 60. Darker colours represent higher divergences, i.e. where the loop multiplicity distribution under RSM deviates the most from the IEA distribution. The y-axes in Figure 7 show the variance differences Δ_{ii} which in Theorem 2 were proven to be less or equal to zero, i.e. $V_{\text{RSM}}(M_{ii}) \leq V_{\text{IEA}}(M_{ii})$. We see that the distribution of loop multiplicity under RSM is more closely related to that of the IEA distribution when $d_i/2m < 0.5$ and Δ_{ii} is close to zero. However, as seen in the case for edge multiplicities in Figure 6, the maximal value of divergence increases when m increases. The divergence increases monotonically from zero to this maximal value, and decreases very steeply back to zero for the degenerate case with the stub proportion 1 and $\Delta_{ii} = 0$ (see Theorem 2).

4.4 A Numerical Example

A small numerical example is given to highlight the results from the theoretical comparison of the local edge multiplicities and the tendencies seen when comparing their simulated distributions. We use the multigraph in Figure 1 and consider local vertex and dyad sites to see how local properties at these sites vary over the network.

Table 1 gives the local structure at dyad $(i, j) \in R$ with (d_i, d_j) local stubs and $2m - d_i - d_j$ the external stubs. The local information here comprise the six multiplicities given by the multiplicity sequence $\mathbf{m}_{ij} = (m_{ii},\ m_{ij},\ m_{jj},\ m_{i*},\ m_{j*},\ m_{**})$ where $*$ indicates a summation over all vertices except i and j so that

$$m_{i*} = \sum_{\substack{u \\ u \neq i,j}} (m_{iu} + m_{ui}), \quad m_{j*} = \sum_{\substack{u \\ u \neq i,j}} (m_{ju} + m_{uj}), \quad m_{**} = \sum_{\substack{u \leq v \\ u \neq i,j \\ v \neq i,j}} m_{uv} \ .$$

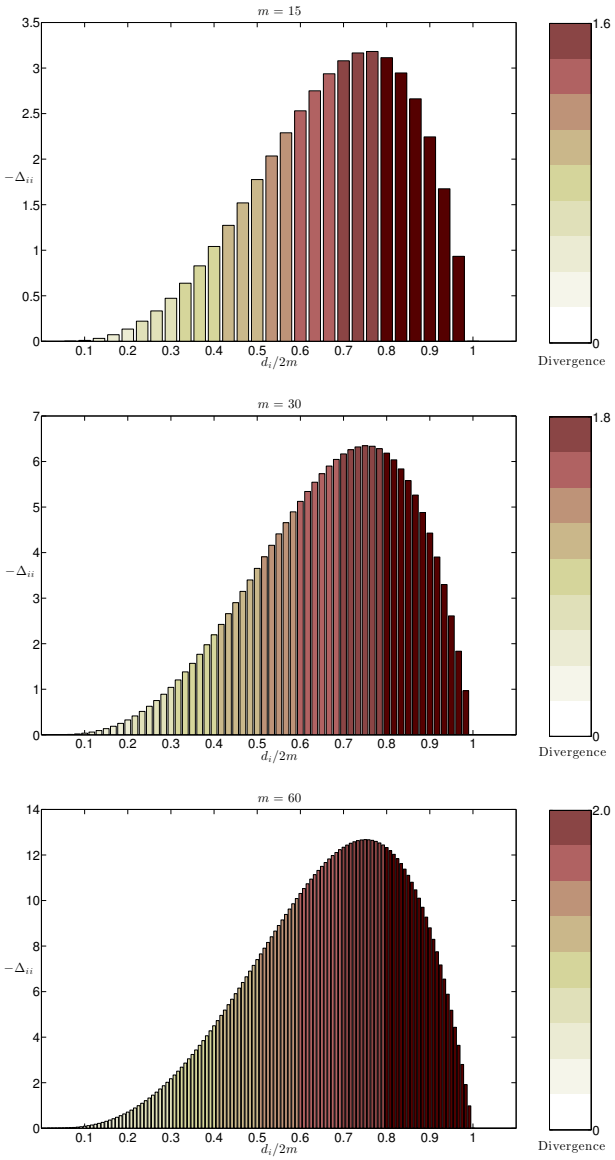

Figure 7: Divergence and negative variance difference $-\Delta_{ii}$ between the multiplicity distribution under random stub matching and under independent edge assignments for loops at a vertex of degree d_i in graphs with m=15, 30 and 60 edges. The variance difference is plotted against stub proportion $d_i/2m$, and a darker colour at the bars located on these proportions represent a larger divergence than a brighter colour.

For each selected dyad we estimate the common expected value $E(M_{ij})$ and the variance difference Δ_{ij} where $i < j$. Further, we calculate the divergence between the simulated distributions of M_{ij} under the two models which indicates the resemblance or discrepancy between the two distributions. The following is noted from Table 1. The two distributions resemble each other the most when there is a large number of external stubs and edge multiplicities at other vertices than the ones included in the selected dyad. This occurs for dyad A and B where the number of external stubs is equal to 25 and the divergence is equal to 0.03. Moreover, it is for this dyad that the variance difference Δ_{ij} is closest to zero.

Table 2 shows the local structure of vertex sites with d_i local stubs, $d_i/2m$ proportion of local stubs, and $2m-d_i$ external stubs. The local information here comprise $(m_{ii},\ m_{i*},\ ,m_{**})$ where, as above, $*$ denotes the summation over all vertices except vertex i. As in Table 1,

we estimate the expected value and variance difference, and calculate the divergence between the loop multiplicity distributions under the two models. Consistent with results seen in Section 4.3 with simulated loop distributions, the lower the stub proportion at vertex i, the more similar are the two distributions. This is seen for vertex A with stub proportion 0.18 and divergence 0.03. Further, the variance difference is closest to zero for this case.

The results and tendencies seen can be summarised as follows. For both loop and non-loop multiplicity distributions, the more external stubs we have at other vertices than the local vertex or dyad under study, the better we can approximate RSM($\mathbf{d}$) with an IEA(m,$\mathbf{Q}(\mathbf{d})$). Note that we have assumed that both the number of vertices n and the number of edges m are known or estimable for the multigraph with unknown multiplicity sequence $\mathbf{m} = (m_{ij} : (i,j) \in R)$.

Table 1: Local structure of dyad sites in the aggregated multigraph of Figure 1 with information about loops and non-loops within the dyad, adjacent other non-loops at the two vertices in the dyad, and external loops and non-loops. The common expected values $E(M_{ij})$ and variance differences Δ_{ij} are estimated under independent edge assignment (IEA) and random stub matching (RSM), and the divergence is calculated between the simulated edge multiplicity distributions under the two models.

Dyad		Local stubs		External stubs	Local multiplicities			External multiplicities			Under IEA and RSM		
i	j	d_i	d_j	$2m - d_i - d_j$	m_{ii}	m_{ij}	m_{jj}	m_{i*}	m_{j*}	m_{**}	$\hat{E}(M_{ij})$	$\hat{\Delta}_{ij}$	Divergence
A	B	7	8	25	1	1	0	4	7	7	1.44	-0.33	0.03
A	C	7	14	19	1	3	2	2	7	5	2.51	-0.70	0.05
A	D	7	11	22	1	1	3	4	4	7	1.98	-0.50	0.04
B	C	8	14	18	0	5	2	3	5	5	2.87	-0.77	0.05
B	D	8	11	21	0	2	3	6	3	6	2.26	-0.57	0.04
C	D	14	11	15	2	2	3	8	3	2	3.95	-0.94	0.04

Table 2: Local structure of vertex sites in the aggregated multigraph of Figure 1 with information about loops and adjacent non-loops, and external edges at vertex i. The common expected values $E(M_{ii})$ and variance differences Δ_{ii} are estimated under independent edge assignment (IEA) and random stub matching (RSM), and the divergence is calculated between the simulated loop multiplicity distributions under the two models.

Vertex	Local stubs	Local stub proportion	Local multiplicity	External multiplicities		Under IEA and RSM		
i	d_i	$d_i/2m$	m_{ii}	m_{i*}	m_{**}	$\hat{E}(M_{ii})$	$\hat{\Delta}_{ii}$	Divergence
A	7	0.18	1	5	14	0.54	-0.13	0.03
B	8	0.20	0	8	12	0.72	-0.20	0.04
C	14	0.35	2	10	8	2.33	-1.01	0.14
D	11	0.28	3	5	12	1.41	-0.52	0.08

5 Global structure of multigraphs

In this section, we analyse the global structure of random multigraphs by measures defined using the distribution of edge multiplicities. More specifically, we show how previous results on the local edge multiplicities can be extended to investigate some structural network features by using the simultaneous distribution of edge multiplicities. We start with a definition:

graphs with larger number of multiple edges and loops are more complex, in contrast to simple graphs without multiple edges and loops (Wasserman and Faust, 1994). Following this definition, we derive moments of statistics that identify complexity and simplicity in multigraphs. The probability distribution of the complexity of multigraphs generated by RSM depends in a complicated way on its degree sequence $\mathbf{d}$. Shafie (2015) derives several complexity statistics and shows how easily handled they are under the IEA(m,$\mathbf{Q}$) model. This implies that these statistics are more easily handled when the stochastic dependence between edge assignments under the RSM($\mathbf{d}$) model can be ignored, so that IEA(m,$\mathbf{Q}(\mathbf{d})$) can be used as an approximation.

Examples of useful information about complexity are given by the expected number of loops and multiple edges, and their variances. In particular, for aggregated multigraphs based on vertex attributes, the number of loops informs on edge moving within categories, thus also indicating the tendency for homophily.

The expected number of loops M_1 and the expected number of non-loops M_2 under IEA(m,$\mathbf{Q}$) are given by

$$E_{\text{IEA}}(M_1) = \sum_{i=1}^{n} E(M_{ii}) = m \sum_{i=1}^{n} Q_{ii} \ , \tag{7}$$

$$E_{\text{IEA}}(M_2) = \sum_{i<j} \sum E(M_{ij}) = m \sum_{i<j} \sum Q_{ij} \ , \tag{8}$$

and their common variance is given by

$$V_{\text{IEA}}(M_1) = \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(M_{ii}, M_{jj}) = m \left[\sum_{i=1}^{n} Q_{ii}(1 - Q_{ii}) - \sum_{i \neq j} \sum Q_{ii} Q_{jj} \right] \ , \tag{9}$$

where Q_{ij} for $i \leq j$ denotes some generic edge assignment probabilities (Shafie, 2015). To summarise the distribution of M_1 and M_2 under the RSM($\mathbf{d}$) model, we give their first two central moments in Theorem 3.

Theorem 3. *Let M_1 and M_2 be the number of loops and non-loops, respectively. For $1 \leq d_i \leq d_j \leq 2m$, their expected values and common variance are*

$$E_{\text{RSM}}(M_1) = \frac{1}{2m-1} \sum_{i=1}^{n} \binom{d_i}{2} \ ,$$

$$E_{\text{RSM}}(M_2) = \frac{1}{2m-1} \sum_{i<j} \sum d_i d_j \ ,$$

and

$$V_{\text{RSM}}(M_1) = \frac{1}{2m-1} \sum_{i=1}^{n} \binom{d_i}{2} \left[1 - \frac{1}{2m-1} \sum_{j=1}^{n} \binom{d_j}{2} \right] + \frac{1}{(2m-1)(2m-3)} \left[\sum_{i=1}^{n} \binom{d_i}{2} \binom{d_i - 2}{2} + \sum_{i \neq j} \sum \binom{d_i}{2} \binom{d_j}{2} \right] \ .$$

Proof. The moments are obtained directly as expected values of the local multiplicities given in Equation (7) to (9) by substituting the edge assignment probabilities Q_{ij} from Equation (2). Note that the common variance is due to the linear relationship $M_2 = m - M_1$. □

Corollary 3. *For regular multigraphs with the same degree d at every vertex $i = 1, \ldots, n$, it follows that*

$$E_{\text{RSM}}(M_1) = \frac{d-1}{2}\left(1 + \frac{1}{nd-1}\right)$$

and

$$V_{\text{RSM}}(M_1) = \frac{d-1}{2}\left(1 + \frac{1}{nd-1} + \frac{(d-2)(d-3)}{2nd}\right) + O\left(\frac{1}{n^2}\right) \ .$$

From Corollary 3 we expect that there are slightly more than $(d-1)/2$ loops, and the expected number of loops is about the same for any number of vertices. Since the variance is approximately equal to the expected value, the number of loops might be approximately Poisson distributed.

The moments of several other statistics given in Shafie (2015) can easily be derived for the RSM model by substituting the edge assignment probability sequence $\mathbf{Q}$ with $\mathbf{Q}(\mathbf{d})$. Some examples are M_3 denoting the number of pairs of equal non-loops given by

$$M_3 = \sum_{i<j}\sum \binom{m_{ij}}{2} = \sum_{i<j}\sum_{k<\ell}\sum\sum I_{ijk} I_{ij\ell} \ ,$$

and the sum $M_1 + M_3$ which is equal to zero if and only if the multigraph is simple. The expected values for these statistics are given by

$$E_{\text{RSM}}(M_3) = \frac{m(m-1)}{2}\sum_{i<j}\sum Q_{ijij} = \frac{2}{(2m-1)(2m-3)}\sum_{i<j}\sum \binom{d_i}{2}\binom{d_j}{2} \ ,$$

and

$$\begin{aligned} E_{\text{RSM}}(M_1 + M_3) &= m\sum_{i=1}^{n} Q_{ii} + \binom{m}{2}\sum_{i<j}\sum Q_{ijij} \\ &= \frac{1}{2m-1}\sum_{i=1}^{n}\binom{d_i}{2} + \frac{2}{(2m-1)(2m-3)}\sum_{i<j}\sum \binom{d_i}{2}\binom{d_j}{2} \ . \end{aligned}$$

In particular, for regular graphs with the same degree d at every vertex, this expected value is about $E_{\text{RSM}}(M_1 + M_3) = (d^2-1)/4$ regardless of the number of vertices. Other statistics for analysing global properties of multigraphs are related to the complexity sequence $\mathbf{R} = (R_0, R_1 \ldots, R_m)$, where

$$R_k = \sum_{i\leq j}\sum I(M_{ij} = k) \qquad \text{for} \quad k = 0, 1, \ldots, m \ ,$$

and counts the number of multiplicities equal to k (Frank and Shafie, 2012). The expected value and variance of R_k are more complicated to derive under RSM. However, Shafie (2015) derives both of these under the IEA model and further describes other summary measures of $\mathbf{R}$ that are of interest for studying the complexity of multigraphs and for detecting structural dependencies. Moreover, Shafie (2015) shows how interval estimates of R_k for $k = 0, 1, \ldots, m$, in combination with interval estimates of M_1 and M_2, can be used to check whether structural dependencies, e.g. tendency for vertex isolation, strengthening existing ties and multiplexity, are present in the empirical network.

We introduce yet another random complexity measure which can be used to determine the probability distribution of multigraphs under RSM($\mathbf{d}$).

Theorem 4. *The probability of specified undirected multigraphs under RSM(***d***) depends on a single non-negative complexity statistic t according to*

$$P(\mathbf{M}=\mathbf{m}) = \frac{\binom{m}{\mathbf{m}}\ 2^{m_2}}{\binom{2m}{\mathbf{d}}} = \frac{m!\ \ 2^m\ d_1!\cdots d_n!}{(2m)!\ \ 2^t}\ ,$$

where

$$t = \sum_{i=1}^{n} m_{ii} + \sum_{i\le j}\sum \log_2 m_{ij}!\ . \tag{10}$$

Proof. The randomly permuted stub sequence $\mathbf{X}$ has

$$\binom{2m}{\mathbf{d}} = \frac{(2m)!}{d_1!\dots d_n!}$$

different outcomes with equal probabilities. Each outcome $\mathbf{m}$ of $\mathbf{M}$ corresponds to

$$\binom{m}{\mathbf{m}} = \frac{m!}{\prod_{i\le j} m_{ij}!}$$

outcomes of $\mathbf{Y}$ and each outcome of $\mathbf{Y}$ corresponds to 2^{m_2} outcomes of $\mathbf{X}$ where

$$m_2 = \sum_{i<j}\sum m_{ij} = m - \sum_{i=1}^{n} m_{ii}\ .$$

It follows that the probability of $\mathbf{M}=\mathbf{m}$ is given by

$$\begin{aligned} P(\mathbf{M}=\mathbf{m}) = \frac{\binom{m}{\mathbf{m}}\ 2^{m_2}}{\binom{2m}{\mathbf{d}}} &= \frac{m!\ \ d_1!\dots d_n!\ \ 2^{m-\sum_{i=1}^n m_{ii}}}{(2m)!\prod_{i\le j} m_{ij}!} \\ &= \frac{m!\ \ d_1!\dots d_n!\ \ 2^m}{(2m)!\,2^{\sum_{i=1}^n m_{ii}}\prod_{i\le j} m_{ij}!} \end{aligned} \tag{11}$$

for all possible outcomes $\mathbf{m} = (m_{ij} : (i,j) \in R)$ satisfying the restrictions imposed by the degree sequence $\mathbf{d}$:

$$\begin{aligned} d_1 &= 2m_{11} + m_{12} + \cdots + m_{1n} \\ d_2 &= m_{12} + 2m_{22} + m_{23} + \cdots + m_{2n} \\ &\vdots \\ d_n &= m_{1n} + \cdots + m_{n-1,n} + 2m_{nn}\ . \end{aligned}$$

Rewriting Equation (10) as

$$2^t = 2^{\sum_{i=1}^n m_{ii}} \prod_{i\le j} m_{ij}!$$

and inserting it into Equation (11) yields the result. □

The statistic t in Equation (10) is a complexity measure that takes on positive values when at least one vertex pair site has loops or multiple edges. Following Theorem 4, we obtain more results concerning the simplicity and complexity of a multigraph, and the probability distribution of the edge multiplicities under RSM($\mathbf{d}$).

Corollary 4. *Outcomes* $\mathbf{m}$ *with the same value* $t = t(\mathbf{m})$ *have the same probability, and these probabilities decrease with increasing values of* t*.*

Corollary 5. *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 expected value for the random complexity statistic T is complicated to specify under RSM($\mathbf{d}$). However, Theorem 5 gives $E_{\text{IEA}}(T)$.

Theorem 5. *The expected value of the random complexity measure*

$$T = \sum_{i=1}^{n} M_{ii} + \sum\sum_{i \le j} \log M_{ij}!$$

under IEA(m*,*$\mathbf{Q}$*) is given by*

$$E_{\text{IEA}}(T) = m \sum_{i=1}^{n} Q_{ii} + \sum_{k=2}^{m} \left[\log k! \binom{m}{k} \sum\sum_{i \le j} Q_{ij}^k (1 - Q_{ij})^{m-k} \right] \ .$$

Proof.

$$\begin{aligned} E_{\text{IEA}}(T) &= E_{\text{IEA}} \left[\sum_{i=1}^{n} M_{ii} + \sum\sum_{i \le j} \log M_{ij}! \right] \\ &= E_{\text{IEA}} \left[M_1 + \sum_{k=2}^{m} R_k \log k! \right] \ , \end{aligned} \tag{12}$$

where the expected value of M_1 is given in Equation (7), and the expected values of R_k is equal to

$$E_{\text{IEA}}(R_k) = \binom{m}{k} \sum\sum_{i \le j} Q_{ij}^k (1 - Q_{ij})^{m-k} \quad \text{for} \quad k = 0, 1, \ldots, m \ ,$$

(Shafie, 2015). Inserting $E_{\text{IEA}}(M_1)$ and $E_{\text{IEA}}(R_1)$ into Equation (12) yields the result.

□

6 Conclusion and future topics

In this article, we investigate the simultaneous (global) distribution of edge multiplicities as well as the marginal distributions of (local) number of loops and non-loops at specified vertices. This is done under two random multigraph models. The first model is obtained by independent edge assignments (IEA) to vertex pair sites according to a common probability distribution over the sites, and is more simple to analyse due to the independence assumption. The second multigraph model is random stub matching (RSM) where the edges are formed by randomly coupling pairs of stubs according to a fixed stub multiplicity or degree sequence. Thus, edge assignments to vertex pair sites are dependent. If we ignore the dependency between edges in the RSM model and assume independent edge assignments of stubs, we obtain an approximate IEA model from an RSM model. This is desirable since many statistics are easily derived under IEA and will thus facilitate the structural analysis (Shafie, 2015). In order to determine when this approximation is suitable to use, we derive some general results under RSM and IEA using the two central moments of the number of loops at a fixed vertex and the number of edges between two distinct vertices. These moments measure the centrality and spread of the distributions and are determined as functions of the number of edges, denoted m, and

the degrees of the vertices. Further, simulations are performed to calculate the information divergence which indicates when there is resemblance and discrepancy between the marginal edge multiplicity distributions under RSM and approximate IEA.

The main results concerning the distributions of edge multiplicities at local sites can be summarised as follows. The variance of the number of loops under RSM is shown to be less than the variance under the approximate IEA, except for some degenerate cases. The variance of the number of edges between two distinct vertices under RSM is generally less than the variance under the approximate IEA, except for special cases where the degrees of the two vertices lie symmetrically around m and are given by $m \pm k$ for any non-negative integer k less than a specified limit. Special attention was paid to these cases since it is only for these that the variance under RSM is greater than that of the approximate IEA.

The simulations show the following tendencies: the edge multiplicity distribution (loop and non-loop) under RSM is closely related to that of the approximate IEA distribution when there is a large number of stubs at other vertices than the vertex or dyad under study. In other words, the distribution of loop and edge multiplicity under RSM is closely related to that of the approximate IEA distribution at vertices with low degrees. Moreover, if the difference of the edge multiplicity variances under RSM and approximate IEA are equal or close to zero, we see more resemblance between the two distributions.

Some extensions to analyse the global structure of multigraphs using the simultaneous or multivariate distribution of edge multiplicities are also given. Different aspects of complexity, defined and quantified by the distribution of edge multiplicities, can be studied by various indicators and summary measures. The moments of some suggested statistics that identify simplicity and complexity are given and shown to be more easily handled under IEA. This further points to the relevance of knowing when an IEA approximation can be used. Further, the distribution of multigraphs under RSM is shown to depend on a single complexity statistic. Multigraphs of the same complexity have equal probabilities which decreases with increasing complexity. Since complexity is related to a single sufficient statistic, the RSM model can be expressed as an exponential random multigraph model (ERMM) (Lusher et al., 2012; Koehly and Pattison, 2005) with complexity as the only statistic in the linear exponent. For an ERMM representation of the IEA model, see Shafie (2015).

The edge multiplicity distributions under IEA and RSM have sometimes very different support sets due to outcomes being restricted by the RSM degree sequence. In particular, it was shown that the range of the loop multiplicity distribution under RSM is only a fraction of that of the IEA distribution. This means that the divergence, which is obtained as a non-zero weighted sum of the likelihood ratios, is calculated for this fraction only and may not be a reliable measure for comparing the two distributions. This points to the need for more general results obtained by a thorough theoretical investigation of, and comparison between, the marginal distributions. This is a topic for future research. Moreover, the results indicate that the multigraph distributions under IEA and RSM have different spreads around common central edge multiplicities which should be investigated further. The Bayesian approximation of the RSM model introduced in Section 4.1 can be analysed and compared to the IEA model in similar fashion.

We end with some other research directions which are worthwhile to consider in the future. It was briefly mentioned in this article how local aggregated multigraphs can be used to analyse global properties of networks. These kinds of aggregated multigraphs are obtained from vertex or dyads sites and include a fictitious vertex collapsing all other vertices excluded from the local site under study. The data may for example consist of a single dyad but data from several dyads selected by simple random sampling can be used to obtain better precision. It is therefore of interest to perform studies where focus is on bias in inference from local to global properties and on relationships between local and global measures of structure.

An important further step of this analysis is to compare empirical networks with the

methods and models presented here. This in order to check if observed data, on the aggregate macro level, exhibit similar or different features than those one would except under the models. Thus, applications will guide in future model specifications and provide insight into apparent phenomena in aggregated social networks.

Acknowledgements

I would like to thank Ove Frank for his insightful comments and valuable suggestions during the preparation of this manuscript. I would also like to thank the anonymous reviewers for their constructive comments that help improve the manuscript. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007- 2013)/ERC Grant Agreement n. 319209.

References

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

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

Bollobás, B. (1980). A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316.

Boorman, S. A. and White, H. C. (1976). Social structure from multiple networks. II. role structures. *American Journal of Sociology*, pages 1384–1446.

Davis, J. A. and Leinhardt, S. (1972). The structure of positive interpersonal relations in small groups. In Berger, J., Zeldith Jr, M., and Anderson, B., editors, *Sociological Theories in Progress, Volume 2*, pages 218–251. Houghton Mifflin.

de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306.

Erdős, P. and Rényi, A. (1959). On random graphs, I. *Publicationes Mathematicae Debrecen*, 6:290–297.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61.

Fienberg, S. E., Meyer, M. M., and Wasserman, S. S. (1985). Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80(389):51–67.

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

Frank, O. (2011). Statistical information tools for multivariate discrete data. In Pardo, L., Balakrishnan, N., and Gil, M. A., editors, *Modern Mathematical Tools and Techniques in Capturing Complexity*, pages 177–190. Springer Berlin Heidelberg.

Frank, O. and Shafie, T. (2012). Complexity of families of multigraphs. In *JSM Proceedings, Section on Statistical Graphics*, pages 2908–2921. Alexandria, VA: American Statistical Association.

Frank, O. and Shafie, T. (2016). Multivariate entropy analysis of network data. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 129(1):45–63.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

Gray, R. M. (2011). *Entropy and Information Theory*. Springer Science & Business Media.

Holland, P. W. and Leinhardt, S. (1970). A method for detecting structure in sociometric data. *American Journal of Sociology*, pages 492–513.

Holland, P. W. and Leinhardt, S. (1976). Local structure in social networks. *Sociological Methodology*, 7(1):1–45.

Jackson, M. O. (2005). A survey of network formation models: stability and efficiency. In Demange, G. and Wooders, M., editors, *Group Formation in Economics: Networks, Clubs, and Coalitions*, pages 11–49. Cambridge University Press.

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

Koehly, L. M. and Pattison, P. (2005). Random graph models for social networks: Multiple relations or multiple raters. In Carrington, P. J., Scott, J., and Wasserman, S., editors, *Models and Methods in Social Network Analysis*, pages 162–191. Cambridge university press.

Kullback, S. (1968). *Information Theory and Statistics*. Courier Corporation.

Lazega, E. and Pattison, P. E. (1999). Multiplexity, generalized exchange and cooperation in organizations: a case study. *Social networks*, 21(1):67–90.

Lusher, D., Koskinen, J., and Robins, G. (2012). *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press.

McKay, B. D. and Wormald, N. C. (1990). Uniform generation of random regular graphs of moderate degree. *Journal of Algorithms*, 11(1):52–67.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, pages 415–444.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Newman, M. E., Strogatz, S. H., and Watts, D. J. (2003). Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):167–256.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic block-structures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Pattison, P. (1993). *Algebraic Models for Social Networks*. Cambridge University Press.

Pattison, P. and Wasserman, S. (1999). Logit models and logistic regressions for social networks: II. multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52(2):169–194.

Pattison, P. and Wasserman, S. (2002). Multivariate random graph distributions: applications to social network analysis. In Hagberg, J., editor, *Contributions to Social Network Analysis, Information Theory and Other Topics in Statistics: A Festschrift in Honour of Ove Frank on the Occasion of His 65th Birthday*, pages 74–100. University of Stockholm.

Pattison, P., Wasserman, S., Robins, G., and Kanfer, A. M. (2000). Statistical evaluation of algebraic constraints for social networks. *Journal of Mathematical Psychology*, 44(4):536–568.

Robins, G. (2013). A tutorial on methods for the modeling and analysis of social network data. *Journal of Mathematical Psychology*, 57(6):261–274.

Shafie, T. (2015). A multigraph approach to social network analysis. *Journal of Social Structure*, 16(1):1–22.

Snijders, T. A. (2011). Statistical models for social networks. *Annual Review of Sociology*, 37:131–153.

Wang, P. (2012). Exponential random graph model extensions: Models for multiple networks and bipartite networks. In Lusher, D., Koskinen, J., and Robins, G., editors, *Exponential Random Graph Models for Social Networks*, pages 115–129. Cambridge University Press.

Wasserman, S. (1987). Conformity of two sociometric relations. *Psychometrika*, 52(1):3–18.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

White, H. C., Boorman, S. A., and Breiger, R. L. (1976). Social structure from multiple networks I. blockmodels of roles and positions. *American Journal of Sociology*, pages 730–780.

Wormald, N. C. (1980). Some problems in the enumeration of labelled graphs. *Bulletin of the Australian Mathematical Society*, 21(1):159–160.

Wormald, N. C. (1981). The asymptotic connectivity of labelled regular graphs. *Journal of Combinatorial Theory, Series B*, 31(2):156–167.

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