



Random Multigraphs

Complexity Measures, Probability Models and Statistical Inference

Termeh Shafie

Doctoral Dissertation
Department of Statistics
Stockholm University
S-106 91 Stockholm
Sweden

Abstract

This thesis is concerned with multigraphs and their complexity which is defined and quantified by the distribution of edge multiplicities. Two random multigraph models are considered. The first model is random stub matching (RSM) where the edges are formed by randomly coupling pairs of stubs according to a fixed stub multiplicity sequence. The second model is obtained by independent edge assignments (IEA) according to a common probability distribution over the edge sites. Two different methods for obtaining an approximate IEA model from an RSM model are also presented.

In Paper I, multigraphs are analyzed with respect to structure and complexity by using entropy and joint information. The main results include formulae for numbers of graphs of different kinds and their complexity. The local and global structure of multigraphs under RSM are analyzed in Paper II. The distribution of multigraphs under RSM is shown to depend on a single complexity statistic. The distributions under RSM and IEA are used for calculations of moments and entropies, and for comparisons by information divergence. The main results include new formulae for local edge probabilities and probability approximation for simplicity of an RSM multigraph. In Paper III, statistical tests of a simple or composite IEA hypothesis are performed using goodness-of-fit measures. The results indicate that even for very small number of edges, the null distributions of the test statistics under IEA have distributions that are well approximated by their asymptotic χ^2 -distributions. Paper IV contains the multigraph algorithms that are used for numerical calculations in Papers I-III.

Keywords: multigraph, vertex labeled graph, edge labeled graph, isomorphism, edge multiplicity, simplicity and complexity, entropy, joint information, information divergence, goodness-of-fit.

© Termeh Shafie
ISBN 978-91-7447-610-1

Printed in Sweden by US-AB, Stockholm 2012
Distributor: Department of Statistics, Stockholm University

To my mother

List of Included Papers

I Complexity of Families of Multigraphs

To appear in *JSM proceedings*, Section on Statistical Graphics, Alexandria, VA: American Statistical Association.

II Random Stub Matching Models of Multigraphs

Research Report 2012:1, Department of Statistics, Stockholm University.

III Statistical Analysis of Multigraphs

Research Report 2012:2, Department of Statistics, Stockholm University.

IV Some Multigraph Algorithms

Research Report 2012:3, Department of Statistics, Stockholm University.

Acknowledgments

This thesis grew out of many twists and turns along the way, and it is with a lot of emotions that I write my *thank you's* to those who stood by my side as I hurdled all the obstacles in its completion.

First and foremost, I offer my most sincere and heartfelt gratitude to Professor Ove Frank. It has been an honor and a privilege to have you as my main supervisor. Words are not adequate to express my appreciation to you but nonetheless, I will try with the few following. Your excellent guidance, enthusiastic encouragement, intellectual humor, and commitment to the highest standards have inspired and motivated me ever since our first meeting. You always managed to boost my confidence in times when I doubted myself and without your willingness to give your time so generously, this thesis would have remained a dream. Thank you for making this a truly rewarding journey despite the work pressure we were both facing.

I am also most grateful to my assistant supervisor, Professor Dan Hedlin, for carefully reading and commenting drafts of this thesis. Your constructive suggestions and ability to identify the discontinuities in my writing continuously challenged my thinking.

In my daily work, I have been blessed with a cheerful group of fellow doctoral students (former and present), who have become true friends over the years. I wish to acknowledge all of you for providing a stimulating and fun environment in which to learn and grow, and for building a support system to help us all stay sane. To my dear friend and old roomie Karin, thank you for being a great listener and more importantly, for being tolerant when I was trapped in my "bubble" not being a good listener. I will forever treasure the fun times we shared in our office which by far was the best and cosiest room in the whole department (*I think we struck gold when we got a bed!*). To Ellinor, thank you for your continuous support and for creating memories to last a life time - we will always have almond butter, Stina and those crazy shopping sprees! A special thanks goes to Feng for always having a moment (*yes, you really are nicer than Cauchy!*) to help me with programming related issues.

I have been fortunate to come across many wonderful friends without whom life would be bleak. Thank you for listening to my complaints in stressful times and for understanding when I was psychically or mentally absent. Most importantly, thank you for making me dance and laugh in times when I needed it the most.

Last but definitely not least, I am grateful to my family for their undivided love and support. To Pedram, thank you for not only being a great and caring older brother, but also a wonderful friend who constantly inspires me to aim higher in life. To my mother Homa, thank you for always encouraging me in all my endeavors and for never leaving me in doubt of your love for me. Your tremendous faith in me could never be justified by logical argument. To you I dedicate this thesis.

Contents

1	Introduction	ii
1.1	Multigraphs and Applications	ii
1.2	Random Multigraph Models	ii
1.3	Entropy and Information Divergence	iii
1.4	Complexity Measures	iv
2	Summary of Papers	v
2.1	Paper I: Complexity of Families of Multigraphs	v
2.2	Paper II: Random Stub Matching Models of Multigraphs	v
2.3	Paper III: Statistical Analysis of Multigraphs	vi
2.4	Paper IV: Some Multigraph Algorithms	vii
	References	viii
	Included Papers	

1 Introduction

1.1 Multigraphs and Applications

Network data involve relational structure representing interactions between actors and are commonly represented by graphs where the actors are referred to as vertices and the relations are referred to as edges connecting pairs of vertices. These kinds of data arise in a variety of fields including computer science, physics, biology, sociology and economics. Statistical analysis of network data is treated in a book by Kolaczyk (2009) and in survey articles by Frank (2005, 2009, 2011b). Many other issues concerning network analysis are also found in the encyclopedia edited by Carrington, Scott and Wasserman (2005), Meyers (2009), and Scott and Carrington (2011).

In this thesis, mainly undirected graphs representing symmetric relations are considered. An edge with both ends connected to a single vertex is called an edge-loop (or shortly loop), and two or more edges connected to the same pair of vertices are called multiple edges. A simple graph is defined as a graph with no loops or multiple edges and a multigraph is defined as a graph where loops and multiple edges are permitted. Multigraphs appear natural in many contexts, for instance social interactions between people during a period of time, business contacts between companies in a region or industry, and internet connections between websites or between email users during a period of time. Multigraphs can also be obtained by different kinds of vertex and edge aggregations. For instance, several simple graphs representing different binary relations can be aggregated to a multigraph. Examples and illustrations of such aggregations are given in Paper III.

1.2 Random Multigraph Models

A random multigraph is a family of multigraphs with a probability distribution, and appropriately chosen it can be a model for a considered application. The degree of a vertex is the number of edges incident to it, with loops counted twice. Various models have been proposed to study random graphs with fixed or modeled degrees, degree distributions or expected degrees. The classical random graph introduced by Erdős and Rényi (1959, 1960) has independent edges and is fully symmetric with a common binomial distribution for the degree at any vertex. The Erdős-Rényi model has been extensively studied but does not address many issues present in real network dynamics. Therefore, several related models have been proposed. Some of these models are briefly reviewed here. A so called small-world model starts with a ring lattice of vertices and a fixed number of edges at each vertex. With some probability p , each edge in the graph is randomly moved to another position according to a procedure called rewiring (Watts and Strogatz, 1998). For p close to 0, the resulting graph is close to regular while for p close to 1, the resulting graph is close to the Erdős-Rényi random graph. In another generalized random graph model, each vertex receives a weight. Given these weights, edges are assigned to sites of vertex pairs indepen-

dently, and the occupation probabilities for different sites are moderated by the weights of the vertices. One such model is the preferential attachment model (Barabási and Albert, 1999) in which the growth of the random graph is modeled by adding edges to the already existing graph in such a way that vertices with large degrees are more likely to be connected to the newly added edges. Several other methods for generating such random graphs can be found in Blitzstein and Diaconis (2011), Bayati, Kim and Saberi (2010), Britton, Deijfen and Martin-Löf (2006), and Chung and Lu (2002).

In this thesis, two main multigraph models are considered. The first model is random stub matching (RSM) which is also referred to as the configuration model or the pairing model by e.g. Janson (2009), Bollobàs (1980), and Bender and Canfield (1978). Stubs or semi-edges are vertices that are paired to an edge. Under RSM, 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. The second multigraph model is obtained by independent edge assignments (IEA) according to a common probability distribution over the sites. Further, two different methods are presented for obtaining an approximate IEA model from an RSM model. The first method is obtained by assuming that the stubs are randomly generated and independently assigned to vertices, called independent stub assignments (ISA), and can be viewed as a Bayesian model for the stub multiplicities under RSM. The second method of obtaining an approximate IEA model is to ignore the dependency between edges in the RSM model and assume independent edge assignments of stubs (IEAS). This can be viewed as repeated assignments with replacements of stubs, whereas RSM is repeated assignments without replacement of stubs.

1.3 Entropy and Information Divergence

Information theoretic tools based on entropy measures can be used to describe, evaluate and compare different models, and they are particularly useful to analyze variability and dependence structures in multivariate data of network type. A survey of these information theoretic tools can be found in Frank (2011a), Gray (2011), and Kullback (1968). The most common units of information are binary digits (bits) that are based on the binary logarithm.

Entropy can intuitively be understood as a measure of information (uncertainty or variability) associated with a random variable. Similarly, joint entropy can be understood as the amount of joint information in two or more random variables. A more technical interpretation of entropy refers to a property of latent codes. Consider repeated independent outcomes of a random variable with N different possible outcomes and with entropy H . The outcomes can be assigned binary sequences of different lengths according to a prefix code that requires in the long run no more than H bits per outcome. This corresponds to 2^H latent code sequences with uniform probabilities instead of N outcomes with arbitrary probabilities. The length of the latent codes, the entropy H , is called the information in the outcomes, and the extra length that a binary code would require for the outcomes,

$\log N - H$, is called the redundancy in the outcomes.

Information divergence compares two distributions with positive probabilities over the same space of N outcomes, $\mathbf{P} = (P_1, \dots, P_N)$ and $\mathbf{Q} = (Q_1, \dots, Q_N)$. In code language, the divergence is the number of additional bits required when encoding a random variable with a distribution \mathbf{P} using an alternative distribution \mathbf{Q} . Thus, the divergence measures the expected number of extra bits required to code samples from \mathbf{P} when using a code based on \mathbf{Q} , rather than using a code based on \mathbf{P} . Formally, the divergence between \mathbf{P} and \mathbf{Q} is given by

$$D(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^N P_i \left[\log \frac{1}{Q_i} - \log \frac{1}{P_i} \right] = \sum_{i=1}^N P_i \log \frac{P_i}{Q_i},$$

which is an expected log-likelihood ratio. With \mathbf{Q} uniform, the divergence equals the redundancy. The divergence is non-negative and zero only when the two distributions are equal.

1.4 Complexity Measures

Complexity is a general property considered in many different contexts and used with or without a specific definition. Complexity in graphs has been given different definitions in the literature. For instance, Karreman (1955) and Mowshowitz (1968) deal with complexity properties of graphs used as models for molecules with chemical bonds between atoms. The complexity concept used in these references is not the same as those in this thesis. However, a common feature of many complexity concepts is that they seem to be well described and analyzed by information measures based on entropy.

In this thesis, the complexity of a multigraph is defined and quantified by the distribution of edge multiplicities, that is the frequencies of vertex pairs with different numbers of multiple edges. Summary measures of this distribution might be of interest as measures of complexity focusing on special properties of the graph. For instance, the proportion of multiple sites or the average multiplicity among multiple sites are simple measures of complexity focusing on any kind of deviation from graphs without multiple edges. If loops are forbidden, this amounts to deviation from graph simplicity. A special class of complexity measures focuses on the frequency of graphs of different kinds that have the same complexity. Since these numbers might be very large, it is convenient to consider logarithmic measures which are similar to measures based on entropy. The problems of judging the complexity of the set of possible multigraphs and of finding distributions of complexity measures in different random multigraphs are considered.

2 Summary of Papers

2.1 Paper I: Complexity of Families of Multigraphs

This paper analyzes multigraphs with or without vertex and edge labels with respect to structure and complexity. Different types of equivalence classes of these multigraphs are considered and basic occupancy models for multigraphs are used to illustrate different graph distributions on isomorphism and complexity. The loss of information caused by ignoring edge and vertex labels is quantified by entropy and joint information. Further, these tools are used for studying random multigraph properties like uncertainty about outcomes, predictability of outcomes, partial complexity measures and flatness of probability distributions. The main findings can be summarized as follows. General formulae for numbers of graphs in equivalence classes of different kinds are derived and compared to entropies. The entropy of random multigraphs is decomposed according to complexity, graph structure, vertex labeling and edge labeling. It is illustrated how complexity can be captured by partial complexity measures and in particular, the loss of information in partial complexity measures is determined. The probability distribution of number of vertex pairs with no or single edges is specified and compared to the probability distribution of total complexity.

2.2 Paper II: Random Stub Matching Models of Multigraphs

The local and global structure of multigraphs under RSM are here analyzed and compared to IEA models using moments, entropies and information divergences. The local structure of the number of loops at a fixed vertex and the number of edges between two distinct vertices are analyzed. Their moments are determined as functions of the number of edges, denoted m , and the degrees of the vertices. Information divergence and entropies are used to compare the marginal edge multiplicity distributions under RSM and IEA. Approximations to the entropies are given and numerically investigated. The main results concerning the distributions of edge multiplicities at local sites can be summarized as follows. The variance of the number of loops under RSM is shown to be less than the variance under 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 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. For these special cases, the entropies are much higher for IEA than for RSM, and entropy approximations are very good for the IEA distributions but not for the RSM distributions. A new formula for the probability of an arbitrary number of loops at a vertex and the more intricate expression for the probability of an arbitrary number of edges at any site is found.

The global structure of multigraphs is analyzed by the multivariate distribution of edge multiplicities. Simplicity and complexity of multigraphs under RSM are investigated. Two well known asymptotic results for the probability that an RSM multigraph is simple are

numerically investigated and an alternative way of approximating this probability is presented. Some other variables that identify simplicity and complexity are proposed and investigated. The main results concerning the global structure of multigraphs can be summarized as follows. The distributions of multigraphs under RSM are shown to depend on a single complexity statistic. Entropies of the RSM and IEA distributions of multigraphs are given and approximate entropies are found using covariance matrices. The exact and approximate entropies are close to the upper bounds of the exact entropies. The multigraph distributions under RSM and IEA are different due to very different ranges. For regular multigraphs, both the RSM and IEA distributions cluster at the high probability sites when more edges are added and are therefore less flat for large values of m . The two asymptotic formulae for the probability that an RSM multigraph is simple do not perform well for multigraphs with small numbers of vertices and edges and the new proposed approximation is shown to perform better. The moments of some suggested variables that identify simplicity and complexity are shown to be more easily handled under IEA, and the ISA model is introduced as a method to get an IEA distribution. Using this method, further approximations to the RSM entropy are derived. For uniform or close to uniform degree distributions, the approximations are good even for small multigraphs, and for skew distributions they are good for multigraphs with many edges. An asymptotic equipartition property is shown to give yet another approximation that works reasonably well except for multigraphs with skew degree sequences and few vertices.

2.3 Paper III: Statistical Analysis of Multigraphs

Statistical properties are here investigated for some probabilistic multigraph models considered in Papers I and II. Multigraph models defined by RSM and the closely related IEA models are statistically analyzed by using the multiplicity sequence \mathbf{m} of an observed multigraph with n vertices and m edges. Two particular kinds of IEA models are investigated, both of which can be considered as approximations to RSM models. Tests are based on \mathbf{m} mostly by considering goodness-of-fit statistics S of Pearson type and T of likelihood ratio type. For IEA models it is well known that for large number of edges, these test statistics have asymptotic χ^2 -distributions. Some problems we want to specifically analyze are how the test statistics behave for small m and compare their behaviour under RSM and IEA. To that end critical regions of the goodness-of-fit statistics with a given significance level α according to their asymptotic distributions are chosen, and answers to questions like the following are searched for. Are the actual significance levels of S and T for small m far from α ? Is the convergence of the cumulative distribution functions of S and T slow or rapid? Does it depend on specific parameters in the models? Can better approximations to the actual distributions be obtained by using information about moments and adjustments of the χ^2 -distributions? Can power approximations be made for S or T for small m ? How is power related to parameters of the models? How can RSM be tested and how does RSM

influence the distributions of the goodness-of-fit statistics? The main results obtained can briefly be summarized as follows. Even for very small m , the null distributions of the test statistics S and T under IEA have distributions that are fairly well approximated by their asymptotic distributions. This holds true for testing simple as well as composite hypotheses with different asymptotic distributions. The influence of RSM on both test statistics is substantial for small number of edges and implies a shift of their distributions towards smaller values compared to what holds true for the null distributions under IEA. Tests of RSM can be made by critical regions for \mathbf{m} , but S and T cannot distinguish RSM from IEA. The non-null distributions of S and T needed for determining power can be approximated by adjusted χ^2 -distributions. It is possible to judge how powers depend on the parameters of the IEA models. More details about significance and powers are reported in the paper with numerous numerical illustrations in plots and tables.

2.4 Paper IV: Some Multigraph Algorithms

In the Papers I–III, there are several illustrations that require developments of algorithms for the numerical calculations. I have written these algorithms myself. I am sure that more efficient algorithms can be developed and even found in the computer science literature for some of the cases.

References

- Barabási, A. L. and Albert, R. (1999), Emergence of Scaling in Random Networks, *Science*, **286**, 509–512.
- Bayati, M., Kim, J.H. and Saberi, A. (2010), A Sequential Algorithm for Generating Random Graphs, *Algorithmica*, **58**, 860–910.
- Bender, E. A. and Canfield, E. R. (1978), The Asymptotic Number of Labeled Graphs with Given Degree Sequences, *Journal of Combinatorial Theory Series A*, **24(3)**, 296–307.
- Blitzstein, J. and Diaconis, P. (2011), A Sequential Importance Sampling Algorithm for Generating Random Graphs with Prescribed Degrees, *Internet Mathematics*, **6(4)**, 489–522.
- 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.
- Britton, T., Deijfen, M. and Martin-Löf A. (2006), Generating Simple Random Graphs with Prescribed Degree Distribution, *Journal of Statistical Physics*, **124(6)**, 1377–1397.
- Carrington, P., Scott, J. and Wasserman, S. (eds.) (2005), *Models and Methods in Social Network Analysis*, New York: Cambridge University Press.
- Chung, F. and Lu, L. (2002), Connected Components in Random Graphs with Given Expected Degree Sequences, *Annals of Combinatorics*, **6**, 125–145.
- Erdős, P. and Renyi, A. (1959), On Random Graphs, *Publicationes Mathematicae*, Debrecen, **6**, 290–297.
- Erdős, P. and Renyi, A. (1960), On the Evolution of Random Graphs, *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, **5**, 17–61.
- Frank, O. (2005), Network Sampling and Model Fitting, in *Models and Methods in Social Network Analysis*, eds. P. Carrington, J. Scott and S. Wasserman, New York: Cambridge University Press, 31–56.
- Frank, O. (2009), Estimation and Sampling in Social Network Analysis, in *Encyclopedia of Complexity and Systems Science*, ed. R. Meyers, New York: Springer Verlag, 8213–8231.
- Frank, O. (2011a), Statistical Information Tools for Multivariate Discrete Data, in *Modern Mathematical Tools and Techniques in Capturing Complexity*, eds. L. Pardo, N. Balakrishnan and M. Ángeles Gil, Berlin: Springer Verlag, 177–190.
- Frank, O. (2011b), Survey Sampling in Networks, in *Handbook of Social Network Analysis*, eds J. Scott and P. Carrington, London: Sage Publications.
- Gray, R. M. (2011), *Entropy And Information Theory*, New York: Springer Verlag.
- Janson, S. (2009), The Probability that a Random Multigraph is Simple, *Combinatorics, Probability and Computing*, **18(1–2)**, 205–225.
- Karreman, G. (1955), Topological Information Content and Chemical Reactions, *Bulletin of Mathematical Biophysics*, **17**, 279–285.
- Kolaczyk, E. (2009), *Statistical Analysis of Network Data*, New York: Springer Verlag.

Kullback, S. (1959), *Information Theory and Statistics*, Wiley, New York.

Meyers, R. (ed.) (2009), *Encyclopedia of Complexity and Systems Science*, New York: Springer Verlag.

Mowshowitz, A. (1968), Entropy and the Complexity of Graphs: I. An Index of the Relative Complexity of a Graph, *Bulletin of Mathematical Biophysics*, **30**, 175–204.

Scott, J. and Carrington, P. (eds.) (2011), *Handbook of Social Network Analysis*, London: Sage Publications.

Watts, D. J and Strogatz, S. H. (1998), Collective Dynamics of ‘Small-World’ Networks, *Nature*, **393**, 440–442.

Complexity of Families of Multigraphs

Ove Frank and Termeh Shafie

Abstract

This article describes families of finite multigraphs with labeled or unlabeled edges and vertices. It shows how size and complexity vary for different types of equivalence classes of graphs defined by ignoring only edge labels or ignoring both edge and vertex labels. Complexity is quantified by the distribution of edge multiplicities, and different complexity measures are discussed. Basic occupancy models for multigraphs are used to illustrate different graph distributions on isomorphism and complexity. The loss of information caused by ignoring edge and vertex labels is quantified by entropy and joint information that provide tools for studying properties of and relations between different graph families.

Keywords: labeled graph, edge multiplicity, complexity measure, entropy, joint information, isomorphism.

1 Introduction

Typical applications of graphs consider sequences of edges associated with vertex pairs. For instance, records of telephone calls, internet connections, money transactions or business contacts during a period of time and their distributions on pairs of individuals, addresses, bank accounts or companies are four such applications. Multigraphs appear natural in many such contexts. A random multigraph is a family of multigraphs with a probability distribution, and appropriately chosen it can be a model for the application. Information theoretic tools can be used to describe, evaluate and compare different models, and they are particularly useful to analyze variability and dependence structures in multivariate data of network type. A survey of such information theoretic tools based on entropy measures is given by Frank(2011a). Statistical analysis of network data is treated in a book by Kolaczyk (2009) and in survey articles by Frank (2005, 2011b). Many other issues concerning network analysis are also found in the encyclopedia edited by Carrington, Scott and Wasserman (2005), Meyers (2009), and Scott and Carrington (2011).

Frank: *Department of Statistics, Stockholm University, S-106 91 Stockholm, ove.frank@stat.su.se*

Shafie: *Department of Statistics, Stockholm University, S-106 91 Stockholm, termeh.shafie@stat.su.se*

This article focuses on basic occupancy models adapted to fit multigraphs. The complexity of a multigraph is defined as its multiplicity distribution, that is the frequencies of vertex pairs with different numbers of multiple edges. The relationships between labeled and unlabeled graphs, isomorphism and complexity are specified in the next section. The numbers of graphs of various types are given and illustrated in Sections 3 and 4. Section 5 describes different complexity measures. Uniform graph models are analyzed and illustrated in Sections 6 to 8. Some other models are presented in Section 9 together with some comments on extensions and references.

2 Basic Concepts and Notation

A finite graph g with n labeled vertices and m labeled edges associates with each edge an ordered or unordered vertex pair. Let $V = \{1, \dots, n\}$ and $E = \{1, \dots, m\}$ be the sets of vertices and edges labeled by integers, and denote by R the set of available sites of vertex pairs for the edges. For directed graphs $R = V^2$ or $R = \{(i, j) \in V^2 : i \neq j\}$ depending on whether or not loops are allowed. For undirected graphs we use the site space $R = \{(i, j) \in V^2 : i \leq j\}$ or $R = \{(i, j) \in V^2 : i < j\}$ and consider (i, j) with $i \leq j$ as a canonical representation for the unordered vertex pair. Let r be the number of sites so that $r = n^2, n(n-1), \binom{n+1}{2}$ or $\binom{n}{2}$ for the cases mentioned. The graph $g : E \rightarrow R$ is an injective map that is represented by an ordered sequence

$$\mathbf{g} = (g_1, \dots, g_m) \in R^m$$

of m sites for the edges, or, equivalently, by an ordered partition

$$\mathbf{M} = (M_{ij} : (i, j) \in R)$$

of r disjoint subsets of edges for the sites. Here

$$M_{ij} = \{k \in E : g_k = (i, j)\} \quad \text{for } (i, j) \in R.$$

Edges at the same site are called multiple edges, and the number of multiple edges at site (i, j) is the multiplicity denoted by m_{ij} for $(i, j) \in R$. We use the notation

$$\mathbf{g} \leftrightarrow \mathbf{M}$$

for the bijection between the two representations of the graph g .

If edges are not distinguished, their labels can be ignored, and order in \mathbf{g} is irrelevant. A representation for the graph with labeled vertices but unlabeled edges is denoted by \mathbf{g}^* and defined by listing the sites in \mathbf{g} in some canonical order such as

$$(1, 1) < (1, 2) < \dots < (1, n) < (2, 1) < (2, 2) < \dots$$

A convenient shorthand notation is

$$\mathbf{g}^* = ((i, j)^{m_{ij}} : (i, j) \in R) .$$

There is a bijection between the unordered site sequence for the edges and the multiplicity sequence for the edges:

$$\mathbf{g}^* \leftrightarrow \mathbf{m} = (m_{ij} : (i, j) \in R) .$$

If both vertex labels and edge labels are ignored, the isomorphic unlabeled graph is represented by \mathbf{G} . The unordered version of \mathbf{M} is an unordered partition \mathbf{M}^* of the edge set into r subsets. The unordered version of the multiplicity sequence \mathbf{m} is an unordered partition \mathbf{m}^* of m into r non-negative integers. There is a bijection between this partition and the sequence of frequencies of sites with multiplicities $0, 1, \dots, m$ given by $\mathbf{r} = (r_0, \dots, r_m)$ where

$$r_k = \sum_{(i,j) \in R} I(m_{ij} = k) \quad \text{for } k = 0, 1, \dots, m .$$

Thus,

$$\mathbf{m}^* \leftrightarrow \mathbf{r} ,$$

and the sequence \mathbf{r} is called the complexity of the graph g . Figure 1 shows a schematic view of bijections and other functional relationships between the various concepts introduced here. The functional relationships comprise canonizing ordering (denoted by $*$), specifying multiplicities $\mathbf{m} = \mathbf{m}(\mathbf{g})$, specifying isomorphism $\mathbf{G} = \mathbf{G}(\mathbf{m})$ which is a function of \mathbf{m} , and specifying complexity $\mathbf{r} = \mathbf{r}(\mathbf{G})$ which is a function of \mathbf{G} . With an abuse of notation we also write $\mathbf{G} = \mathbf{G}(\mathbf{m}) = \mathbf{G}(\mathbf{g})$ and $\mathbf{r} = \mathbf{r}(\mathbf{G}) = \mathbf{r}(\mathbf{m}) = \mathbf{r}(\mathbf{g})$.

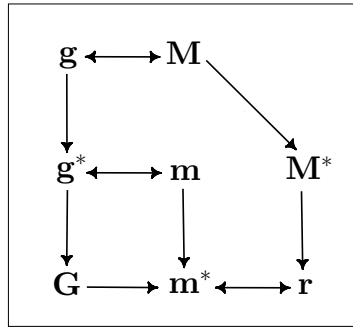


Figure 1: Relationships between graphs, multiplicity and complexity.

3 A Numerical Example

The different concepts introduced are illustrated and visualized by considering a simple example before we turn to general formulae for numbers of graphs and equivalence classes of different kinds.

Consider undirected graphs with $n = 4$ labeled vertices and $m = 3$ labeled edges with loops not allowed so that $r = 6$. Here $m < r$ and all partitions of 3 into positive integers can be used to find the possible multiplicity sequences. Thus the multiplicity sequences divide into three equivalence classes corresponding to permutations of $(3, 0, 0, 0, 0, 0)$, of $(2, 1, 0, 0, 0, 0)$, and of $(1, 1, 1, 0, 0, 0)$. In a shorthand notation, these permutations may be written as $\sim 30^5$, $\sim 210^4$, and $\sim 1^30^3$. The classes have complexity sequences $(5, 0, 0, 1)$, $(4, 1, 1, 0)$, and $(3, 3, 0, 0)$. The classes consist of 1, 2, and 3 non-isomorphic graphs shown in Figure 2.

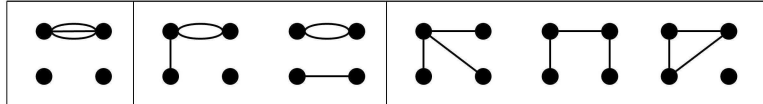


Figure 2: Unlabeled graphs according to complexity.

Each of the non-isomorphic graphs comprise different numbers of vertex and edge labeled graphs. Vertex labels can be assigned to the non-isomorphic graphs in 6, 24, 6, 4, 12, and 4 ways, and edge labels can be assigned to each vertex labeled graph with the same complexity in 1, 3, and 6 ways in the order shown in Figure 2. Table 1 lists the number of unlabeled graphs $\#(\mathbf{G}|\mathbf{r})$, the number of vertex labeled graphs $\#(\mathbf{m}|\mathbf{r})$, and the number of fully labeled graphs $\#(\mathbf{g}|\mathbf{r})$ for each complexity sequence \mathbf{r} . Table 2 gives the numbers $\#(\mathbf{m}|\mathbf{G})$ and $\#(\mathbf{g}|\mathbf{G})$ of vertex labeled and fully labeled graphs for each isomorphism class.

Table 1: Distributions on complexity for graphs with 4 vertices, 3 edges and no loops.

Complexity	(5,0,0,1)	(4,1,1,0)	(3,3,0,0)	Total
Unlabeled graphs	1	2	3	6
Vertex labeled graphs	6	30	20	56
Fully labeled graphs	6	90	120	216

Table 2: Distributions on isomorphism for graphs with 4 vertices, 3 edges and no loops.

Isomorphism							Total
Vertex labeled graphs	6	24	6	4	12	4	56
Fully labeled graphs	6	72	18	24	72	24	216

4 Numbers of Graphs

The number of multigraphs with n labeled vertices, m labeled edges and r available sites of vertex pairs for the edges is given by the number of sequences \mathbf{g} , which is denoted $\#(\mathbf{g}) = r^m$. When edge labels are ignored, the number of graphs is given by the number of multiplicity sequences \mathbf{m} , which is the number of ordered partitions of m into r non-negative integers:

$$\#(\mathbf{m}) = \binom{m+r-1}{m}.$$

Each graph with only vertex labels can be edge labeled in the number of ways that \mathbf{g}^* can be permuted, which is equal to

$$\#(\mathbf{g}|\mathbf{m}) = \binom{m}{\mathbf{m}} = \frac{m!}{\prod_{(i,j) \in R} m_{ij}!}.$$

The total $r^m = \sum_{\mathbf{m}} \binom{m}{\mathbf{m}}$ is a sum over $\binom{m+r-1}{m}$ terms.

Different fully labeled graphs are isomorphic if they are equal when vertex labels as well as edge labels are ignored. The number of isomorphic fully labeled graphs is given by $\binom{m}{\mathbf{m}}$ multiplied by the number of isomorphic vertex labeled graphs with no edge labels. Formally,

$$\#(\mathbf{g}|\mathbf{G}) = \binom{m}{\mathbf{m}} \#(\mathbf{m}|\mathbf{G})$$

since $\binom{m}{\mathbf{m}}$ is invariant for graphs isomorphic to \mathbf{G} .

Multiplicity sequences have the same complexity if they are permutations of the same \mathbf{m}^* . There are $\binom{r}{\mathbf{r}}$ such permutations where $\mathbf{r} \leftrightarrow \mathbf{m}^*$. Thus,

$$\#(\mathbf{m}|\mathbf{r}) = \binom{r}{\mathbf{r}} = \frac{r!}{\prod_{k=0}^m r_k!}$$

is the number of graphs with vertex labels but no edge labels having complexity \mathbf{r} . The number of fully labeled graphs with complexity \mathbf{r} is obtained by multiplying $\#(\mathbf{m}|\mathbf{r})$ with $\#(\mathbf{g}|\mathbf{m})$:

$$\#(\mathbf{g}|\mathbf{r}) = \binom{m}{\mathbf{m}} \binom{r}{\mathbf{r}} = \frac{m! r!}{\prod_{k=0}^m k!^{r_k} r_k!}.$$

The number of different complexity sequences \mathbf{r} is the same as the number of unordered partitions of m into r non-negative integers. This number is the sum of the numbers of unordered partitions of m into k positive integers for $k = 1, 2, \dots, \min(r, m)$. If a_{mk} denotes the number of partitions of m into k positive integers and $a_m = a_{m1} + \dots + a_{mm}$ is the number of partitions of m , it is possible to show that $a_{mk} = a_{m-k}$ for $k \geq m/2$. Tables of a_m and a_{mk} for $k < m/2$ and $m = 1, 2, \dots$ can be used to find

$$\#(\mathbf{r}) = \begin{cases} \sum_{k=1}^r a_{mk} & \text{for } r \leq \frac{m}{2} \\ \sum_{k < m/2} (a_k + a_{mk}) & \text{for } \frac{m}{2} < r < m \\ a_m & \text{for } r \geq m . \end{cases}$$

Such tables are available, for instance, in Comtet (1974).

5 Complexity of Graphs

The complexity sequence \mathbf{r} contains the distribution of multiplicities among the sites. Summary measures of this distribution is of interest as measures of complexity focusing on special properties of the graph. For instance, the proportion of multiple sites

$$\frac{(r - r_0 - r_1)}{r}$$

or the average multiplicity among multiple sites

$$\frac{(m - r_1)}{r - r_0 - r_1}$$

are simple measures of complexity focusing on any kind of deviation from graphs without multiple edges. If loops are forbidden, this amounts to deviation from graph simplicity.

A measure that linearly combines the frequencies of different multiplicities is given by

$$\sum_{k=2}^m \binom{k}{2} r_k ,$$

which counts the number of pairs of edges associated with the same site. If loops are forbidden, this measure is positive if and only if the graph is not simple. Another linear measure with this property is

$$\sum_{k=0}^m r_k \log k! ,$$

which is the logarithm of the common number of permutations that leave the edge sequence \mathbf{g} invariant for graphs of complexity \mathbf{r} .

A special class of complexity measures focuses on how many graphs of different kinds that have the same complexity \mathbf{r} . Since these numbers can be very large, it is convenient to consider logarithmic measures. For vertex labeled and fully labeled graphs with the same complexity, the measures are

$$\log \#(\mathbf{m}|\mathbf{r}) = \log r! - \log \mathbf{r}! = \log r! - \sum_{k=0}^m \log r_k!$$

and

$$\log \#(\mathbf{g}|\mathbf{r}) = \log m! + \log r! - \sum_{k=0}^m (r_k \log k! + \log r_k!) .$$

These measures are similar to measures based on entropy, which characterizes flatness of the relative frequency distributions \mathbf{m}/m and \mathbf{r}/r . Entropy can be considered as a measure of the range or dimension of a latent flat distribution (see Section 7). The entropy of the relative edge frequencies at different sites is given by

$$h(\mathbf{m}/m) = \sum_{(i,j) \in R} \varphi(m_{ij}/m)$$

where

$$\varphi(p) = \begin{cases} -p \log p & \text{for } p > 0 \\ 0 & \text{for } p = 0 . \end{cases}$$

The entropy of the relative site frequencies for different multiplicities is given by

$$h(\mathbf{r}/r) = \sum_{k=0}^m \varphi(r_k/r) .$$

It follows that the entropy of \mathbf{m}/m is equal to

$$h(\mathbf{m}/m) = \log m - \frac{1}{m} \sum_{k=2}^m r_k k \log k .$$

This entropy is non-negative, zero only for all edges at the same site, and it attains a maximal value of $\log r$ only if m is a multiple of r and the multiplicity distribution is uniform with the same multiplicity m/r at all sites. For $m < r$, the maximum is $\log m$ and is attained for all edges at different sites. For other cases with $m > r$ the maximal value is somewhat below $\log r$ and attained for an almost uniform distribution.

It also follows that the entropy of \mathbf{r}/r is equal to

$$h(\mathbf{r}/r) = \log r - \frac{1}{r} \sum_{k=0}^m r_k \log r_k .$$

This entropy is non-negative, zero only for all multiplicities equal, and it attains a maximal value of $\log(m+1)$ only in the degenerate case $m=1, r=2$. The maximal values for other cases are lower but not easily found.

For large values of m , Stirling's formula can be used to show that

$$h(\mathbf{m}/m) = \frac{1}{m} \log \binom{m}{\mathbf{m}} + O\left(\frac{\log m}{m}\right) \approx \frac{1}{m} \log \#(\mathbf{g}|\mathbf{m})$$

so that the entropy of \mathbf{m}/m is approximately equal to the average number of bits (provided logarithms to base 2 are used) per edge needed to generate all \mathbf{g} corresponding to \mathbf{m} .

6 Uniform Graph Models

The classical occupancy models with equal or unequal objects distributed among equal or unequal sites can be modified to fit graph data with its special combinatorial structure for the sites. We focus here on uniform distributions for different families of graphs. Families of graphs are conveniently specified as random graphs. The uniform distributions might be null models used to test or explore empirical graph families. The range of applications for such null models is conveniently extended to families of subgraphs induced by vertices of special kinds.

Assume that $\boldsymbol{\xi}$ is the edge sequence of a random graph that is uniform with probabilities

$$P(\boldsymbol{\xi} = \mathbf{g}) = \frac{1}{r^m} \quad \text{for } \mathbf{g} \in R^m .$$

In this case the probability distributions of the different functions $\mathbf{m}(\boldsymbol{\xi})$, $\mathbf{G}(\boldsymbol{\xi})$, and $\mathbf{r}(\boldsymbol{\xi})$ are simply given as the relative frequencies of outcomes of $\boldsymbol{\xi}$ that are consistent with the outcomes of the functions. Thus

$$\begin{aligned} P(\mathbf{m}(\boldsymbol{\xi}) = \mathbf{m}) &= \frac{\binom{m}{\mathbf{m}}}{r^m} , \\ P(\mathbf{G}(\boldsymbol{\xi}) = \mathbf{G}) &= \frac{\#(\mathbf{g}|\mathbf{G})}{r^m} , \\ P(\mathbf{r}(\boldsymbol{\xi}) = \mathbf{r}) &= \frac{m! r!}{r^m \prod_{k=0}^m k!^{r_k} r_k!} . \end{aligned}$$

The entropy of a random variable is the same as the entropy of its probability distribution, so

$$H(\boldsymbol{\xi}) = \sum_{\mathbf{g}} \varphi(P(\boldsymbol{\xi} = \mathbf{g})) = m \log r .$$

Using calculation rules for entropy (given for instance in Frank, 2011a) it follows that

$$\begin{aligned} H(\mathbf{m}(\boldsymbol{\xi})) &= H(\boldsymbol{\xi}) - E \left[\log \binom{m}{\mathbf{m}(\boldsymbol{\xi})} \right] , \\ H(\mathbf{G}(\boldsymbol{\xi})) &= H(\boldsymbol{\xi}) - E [\#(\mathbf{g}|\mathbf{G}(\boldsymbol{\xi}))] , \\ H(\mathbf{r}(\boldsymbol{\xi})) &= H(\boldsymbol{\xi}) - E \left[\log \binom{m}{\mathbf{m}(\boldsymbol{\xi})} \right] - E \left[\log \binom{r}{\mathbf{r}(\boldsymbol{\xi})} \right] . \end{aligned}$$

Using that $m_{ij}(\boldsymbol{\xi})$ is binomially distributed with parameters m and $1/r$, the entropy of the multiplicity sequence can be expressed as

$$H(\mathbf{m}(\boldsymbol{\xi})) = m \log r - \log m! + r \sum_{k=2}^m \binom{m}{k} \left(\frac{1}{r}\right)^k \left(1 - \frac{1}{r}\right)^{m-k} \log k! .$$

The entropies of $\mathbf{G}(\boldsymbol{\xi})$ and $\mathbf{r}(\boldsymbol{\xi})$ can be numerically evaluated but seem to have no explicit formulae.

Consider now an alternative model with edge sequence $\boldsymbol{\eta}$ assuming that $\mathbf{m}(\boldsymbol{\eta})$ is uniform and that $\boldsymbol{\eta}$ conditional on $\mathbf{m}(\boldsymbol{\eta})$ is uniform. In this case

$$\begin{aligned} P(\mathbf{m}(\boldsymbol{\eta}) = \mathbf{m}) &= \frac{1}{\binom{m+r-1}{m}} , \\ P(\mathbf{r}(\boldsymbol{\eta}) = \mathbf{r}) &= \frac{\binom{r}{\mathbf{r}}}{\binom{m+r-1}{m}} , \\ P(\boldsymbol{\eta} = \mathbf{g}) &= \frac{1}{\binom{m}{\mathbf{m}(\mathbf{g})} \binom{m+r-1}{m}} , \end{aligned}$$

and

$$\begin{aligned} H(\mathbf{m}(\boldsymbol{\eta})) &= \log \binom{m+r-1}{m} , \\ H(\mathbf{r}(\boldsymbol{\eta})) &= \frac{1}{\binom{m+r-1}{m}} \sum_{\mathbf{r}} \binom{r}{\mathbf{r}} \log \binom{r}{\mathbf{r}} - \log \binom{m+r-1}{m} , \\ H(\boldsymbol{\eta}) &= \frac{1}{\binom{m+r-1}{m}} \sum_{\mathbf{m}} \log \binom{m}{\mathbf{m}} + \log \binom{m+r-1}{m} . \end{aligned}$$

The two uniform models considered are in physics referred to as the Maxwell-Boltzmann model with uniform distribution of unequal particles in unequal cells, and the Bose-Einstein model with uniform distribution of equal particles in unequal cells.

For the fully labeled graphs the entropy of ξ is maximal, and the entropy of η deviates by

$$D_1 = H(\xi) - H(\eta)$$

from it. For the vertex labeled graphs the entropy of $\mathbf{m}(\eta)$ is maximal and the entropy of $\mathbf{m}(\xi)$ deviates by

$$D_2 = H(\mathbf{m}(\eta)) - H(\mathbf{m}(\xi))$$

from it. Therefore the reductions in entropy caused by omitting edge labels is larger for ξ than for η ,

$$H(\xi) - H(\mathbf{m}(\xi)) \geq H(\eta) - H(\mathbf{m}(\eta)) ,$$

and the difference between the reductions is equal to the sum $D_1 + D_2$ of the two deviations from maximal entropy. This can also be expressed as the following ordering of the entropies

$$H(\mathbf{m}(\xi)) \leq H(\mathbf{m}(\eta)) \leq H(\eta) \leq H(\xi) .$$

Some of the simplified complexity measures mentioned in Section 5 rely on the frequencies of sites with no or single occupancy only. The distributions of $r_0(\xi)$ and $r_1(\xi)$ are obtained as marginal distributions of $\mathbf{r}(\xi)$. For $r_0(\xi)$ the marginal probabilities are given by

$$\begin{aligned} P(r_0(\xi) = r_0) &= \frac{r!}{r_0! r^m} \sum S_m(r_1, \dots, r_m) \\ &= \frac{r! S(m, r - r_0)}{r_0! r^m} \quad \text{for } r_0 = 0, 1, \dots, r - 1 . \end{aligned}$$

Here the sum extends over (r_1, \dots, r_m) satisfying $\sum_{k=1}^m r_k = r - r_0$ and $\sum_{k=1}^m k r_k = m$. The term

$$S_m(r_1, \dots, r_m) = \frac{m!}{\prod_{k=1}^m k!^{r_k} r_k!}$$

counts the number of partitions of the edge set into r_1 singletons, r_2 parts of size 2, etc. The sum is equal to $S(m, r - r_0)$, which is a Stirling number of the second kind for the number of partitions of an m -set into $r - r_0$ non-empty disjoint subsets.

For the bivariate distribution of $(r_0(\xi), r_1(\xi))$ the probabilities for $r_0 < r$ and $r_1 \leq \min(m, r - r_0)$ are obtained as

$$P(r_0(\xi) = r_0, r_1(\xi) = r_1) = \frac{r!}{r_0! r_1! r^m} \sum S_m(0, r_2, \dots, r_m) ,$$

where the sum extends over (r_2, \dots, r_m) satisfying $\sum_{k=2}^m r_k = r - r_0 - r_1$ and $\sum_{k=2}^m k r_k = m - r_1$. Thus, to evaluate the sum we need to specify the partitions of $m - r_1$ into $r' =$

$r - r_0 - r_1$ integers larger than 2 and find the terms separately. The number of terms is the same as the number of partitions of $m' = m - r_1 - r' = m - r + r_0$ into r' positive integers, that is the number $a_{m'r'}$ given at the end of Section 4.

Upper and lower bounds to the bivariate probability can be found much easier, and they are based on that

$$S_m(0, r_2, \dots, r_m) = \frac{m! S_{m'}(r_2, \dots, r_m)}{m'! \prod_{k=2}^m k^{r_k}},$$

where

$$m' = \sum_{k=1}^{m-1} k r_{k+1} = \sum_{k=2}^m (k-1)r_k = m - r + r_0.$$

Moreover, the geometric mean of the multiplicities is bounded between 2 and the arithmetic mean, which implies that there are bounds α and β so that

$$\alpha = 2^{r'} \leq \prod_{k=2}^m k^{r_k} \leq [(m' + r')/r']^{r'} = \beta$$

for $r' > 0$ and $m' > 0$. Therefore,

$$\frac{m! S_{m'}(r_2, \dots, r_m)}{m'! \beta} \leq S_m(0, r_2, \dots, r_m) \leq \frac{m! S_{m'}(r_2, \dots, r_m)}{m'! \alpha}$$

and, consequently,

$$\frac{r! m! S(m', r')}{r_0! r_1! m'! r^m \beta} \leq P(r_0(\boldsymbol{\xi}) = r_0, r_1(\boldsymbol{\xi}) = r_1) \leq \frac{r! m! S(m', r')}{r_0! r_1! m'! r^m \alpha},$$

where the lower bound to the probability is often quite accurate.

The probability that there are no multiple edges is given by

$$P(r_1(\boldsymbol{\xi}) = m) = P(r_0(\boldsymbol{\xi}) = r - m, r_1(\boldsymbol{\xi}) = m) = \frac{m! \binom{r}{m}}{r^m} \quad \text{for } r \geq m.$$

The distributions of $r_0(\boldsymbol{\eta})$ and $r_1(\boldsymbol{\eta})$ for the model with uniform vertex labeled graphs are given by

$$P(r_0(\boldsymbol{\eta}) = r_0) = \frac{\binom{r}{r_0} \binom{m-1}{r-r_0-1}}{\binom{m+r-1}{m}} \quad \text{for } r_0 = 0, 1, \dots, r-1$$

and

$$P(r_0(\boldsymbol{\eta}) = r_0, r_1(\boldsymbol{\eta}) = r_1) = \frac{\binom{r}{r_0} \binom{r-r_0}{r_1} \binom{m-r+r_0-1}{r-r_0-r_1-1}}{\binom{m+r-1}{m}},$$

for $r_0 < r$ and $r_1 \leq \min(m, r - r_0)$. The first case is proved by noticing that when the r_0 empty sites have been chosen, the remaining $r' = r - r_0$ sites should have at least one edge

per site, and the remaining $m' = m - r'$ edges can be distributed in any of $\binom{m'+r'-1}{m'}$ ways. Similarly, in the second case, when the r_0 empty sites and the r_1 single occupancy sites have been chosen, the remaining $r' = r - r_0 - r_1$ sites should have at least two edges per site, and the remaining $m' = m - r_1 - 2r'$ edges can be distributed in any of $\binom{m'+r'-1}{m'}$ ways.

In this model the probability that there are no multiple edges is given by

$$P(r_1(\boldsymbol{\eta}) = m) = P(r_0(\boldsymbol{\eta}) = r - m, r_1(\boldsymbol{\eta}) = m) = \frac{\binom{r}{m}}{\binom{m+r-1}{m}} \quad \text{for } r \geq m .$$

Now $\binom{m+r-1}{m} = r(r+1)\cdots(r+m-1)/m! \geq r^m/m!$, so obviously the graph property of having no multiple edges has a smaller probability under the $\boldsymbol{\eta}$ -model than under the $\boldsymbol{\xi}$ -model.

The entropy of $(r_0(\boldsymbol{\eta}), r_1(\boldsymbol{\eta}))$ is smaller than the entropy of the complete complexity sequence $\mathbf{r}(\boldsymbol{\eta})$. The difference reveals how much information is lost by using the simpler complexity measure. A simple illustration showing that simple complexity summaries can be quite satisfactory is given in Table 3. We see that the outcomes of (r_0, r_1) match \mathbf{r} quite well, so that there is not much uncertainty about \mathbf{r} when (r_0, r_1) is known. In fact, for the $\boldsymbol{\xi}$ -model the entropies are $H(\mathbf{r}(\boldsymbol{\xi})) = 2.82$ and $H(r_0(\boldsymbol{\xi}), r_1(\boldsymbol{\xi})) = 2.78$ so only about one percent of the information about complexity is lost by using the simpler complexity measure. The univariate entropies $H(r_0(\boldsymbol{\xi})) = 1.95$ and $H(r_1(\boldsymbol{\xi})) = 2.50$ are also retaining almost the same information as the bivariate entropy. For the $\boldsymbol{\eta}$ -model we find $H(\mathbf{r}(\boldsymbol{\eta})) = 3.63$, $H(r_0(\boldsymbol{\eta}), r_1(\boldsymbol{\eta})) = 3.38$, $H(r_0(\boldsymbol{\eta})) = 2.11$, and $H(r_1(\boldsymbol{\eta})) = 2.50$ which implies that about 7% of the information about complexity is lost by the simpler measure.

Table 3: Number of outcomes of complexity $\mathbf{r} = (r_0, r_1, \dots, r_m)$ for given numbers r_0, r_1 of empty and single occupancy sites in graphs with 5 vertices, 8 edges and no loops.

r_0	r_1								
	0	1	2	3	4	5	6	7	8
2									1
3							1		
4					1	1			
5			1	1	1				
6	1	1	2	1					
7	2	2	1						
8	3	1							
9	1								

7 Entropy and Joint Information

Entropies are convenient measures of variation for general random variables. They are also useful to determine dependence and other relationships between several random variables. This can be intuitively understood by considering entropy as a measure of information, and interpreting it as the number of informative binary dimensions in a bijective representation of the outcomes. The technical interpretation of entropy as information refers to a property of latent codes. It is known that repeated independent outcomes of a random variable with N different possible outcomes and entropy H can be assigned binary sequences of different lengths according to a prefix code that requires in the long run no more than H binary digits (bits) per outcome. This corresponds to 2^H latent code sequences with uniform probabilities instead of N outcomes with arbitrary probabilities. The length of the latent codes, the entropy H , is called the information in the outcomes, and any extra length in a binary code for the outcomes, $\log N - H$, is called the redundancy in the outcomes.

When two random variables ξ and η have common bits in their latent codes, they are related, and this relationship is measured by the joint information or joint entropy

$$J(\xi, \eta) = H(\xi) + H(\eta) - H(\xi, \eta) .$$

Joint information is zero if and only if the variables are independent and thus do not reveal any information about each other. Joint information is maximal when one of the variables is completely determined by the other. Joint information is an alternative to correlation and other measures that require numerical variables and specify linear or special non-linear regression relationships. Arbitrary functional relationships as well as various conditional dependence structures can be specified by different combinations of entropy measures. See Frank (2011a) for further details about such possibilities.

The total information in (ξ, η) minus the information in η is the expected remaining information in ξ when η is provided,

$$H(\xi, \eta) - H(\eta) = E[H(\xi|\eta)] ,$$

and the joint information is equal to the original information minus the remaining information in any of the variables according to

$$J(\xi, \eta) = H(\xi) - E[H(\xi|\eta)] = H(\eta) - E[H(\eta|\xi)] .$$

If η is determined by ξ , the difference $H(\xi) - H(\eta)$ is equal to the remaining information in ξ when η is provided, or, in other words, the information in ξ that is lost if nothing more than η is released.

Consider the edge sequence ξ of a random multigraph. The entropy of $\mathbf{G}(\xi)$ measures variation from uniformity or flatness in the probability distribution over the unlabeled

graphs. The joint information of the multiplicity sequence $\mathbf{m}(\boldsymbol{\xi})$ and the complexity sequence $\mathbf{r}(\boldsymbol{\xi})$ is trivially equal to the entropy of $\mathbf{r}(\boldsymbol{\xi})$ because complexity is determined by multiplicity. Less transparent relationships between network properties might be between number of loops and number of multiple sites or any other network characteristics of special interest for the applications. Joint entropies reveal such relationships. Sometimes it is possible to give explicit expressions for the measures. Some examples are given in Section 9.

8 Some Further Illustrations

Numerical algorithms have been developed to handle distributions of edges among vertex pairs for arbitrary values of m and n . Here these algorithms are used for the case with $n = 6$ and $m = 4$ in order to visualize how families of multigraphs are composed of isomorphisms of varying complexity. We also illustrate the distributions on isomorphism and complexity of fully labeled and vertex labeled graphs. We evaluate the possibilities to gain information about them by using partial information.

Table 4 shows complexity distributions for uniform distributions over the fully labeled graphs, over the vertex labeled graphs, and over the unlabeled graphs. Random variables generating these graph families are the earlier defined edge sequences $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. We also consider an edge sequence $\boldsymbol{\zeta}$ with uniform distribution of $\mathbf{G}(\boldsymbol{\zeta})$ over the isomorphisms and with $\boldsymbol{\zeta}$ uniform conditional on $\mathbf{G}(\boldsymbol{\zeta})$.

Table 4: Distributions on complexity for graphs with 6 vertices, 4 edges and no loops.

Complexity	(14,0,0,0,1)	(13,1,0,1,0)	(13,0,2,0,0)	(12,2,1,0,0)	(11,4,0,0,0)	Total
Unlabeled graphs	1	2	2	7	9	21
Vertex labeled graphs	15	210	105	1365	1365	3060
Fully labeled graphs	15	840	630	16380	32760	50625

The complexity distributions have entropies $H(\mathbf{r}(\boldsymbol{\xi})) = 1.11$, $H(\mathbf{r}(\boldsymbol{\eta})) = 1.51$ and $H(\mathbf{r}(\boldsymbol{\zeta})) = 1.91$. Maximal entropy is here equal to $\log 5 = 2.32$. Thus, the complexity distributions have redundancies of 52%, 35%, and 18%, and all distributions exhibit a clear concentration towards simplicity with no or few multiple edges.

From the distributions on isomorphisms in Figure 3 it follows that $H(\mathbf{G}(\boldsymbol{\xi})) = 3.87$, $H(\mathbf{G}(\boldsymbol{\eta})) = 3.95$, and $H(\mathbf{G}(\boldsymbol{\zeta})) = 4.39$. The unlabeled graphs have maximal entropy $\log 21 = 4.39$ obtained for the $\boldsymbol{\zeta}$ -model. The vertex labeled graphs have maximal entropy for the $\boldsymbol{\eta}$ -model, and 34% of that entropy is retained by $\mathbf{G}(\boldsymbol{\eta})$. The fully labeled graphs have maximal entropy for the $\boldsymbol{\xi}$ -model, and 25% of that entropy is retained by $\mathbf{G}(\boldsymbol{\xi})$. A more complete and systematic view of how the information content in different kinds of data varies for the three models is given in Table 5. The models are constructed to have no redundancy

for one of the data levels. All other redundancies are between 4% and 12%, except for the complexity level which has higher redundancies. If maximal entropy is rounded upwards, a rough common feature of the models is apparent. Of the 16 binary dimensions required for fully labeled graphs, about 3 are informative about complexity, another 2 are informative about how the sites need to be ordered to achieve graph structure, another 7 are informative about vertex labeling, and another 4 about edge labeling.

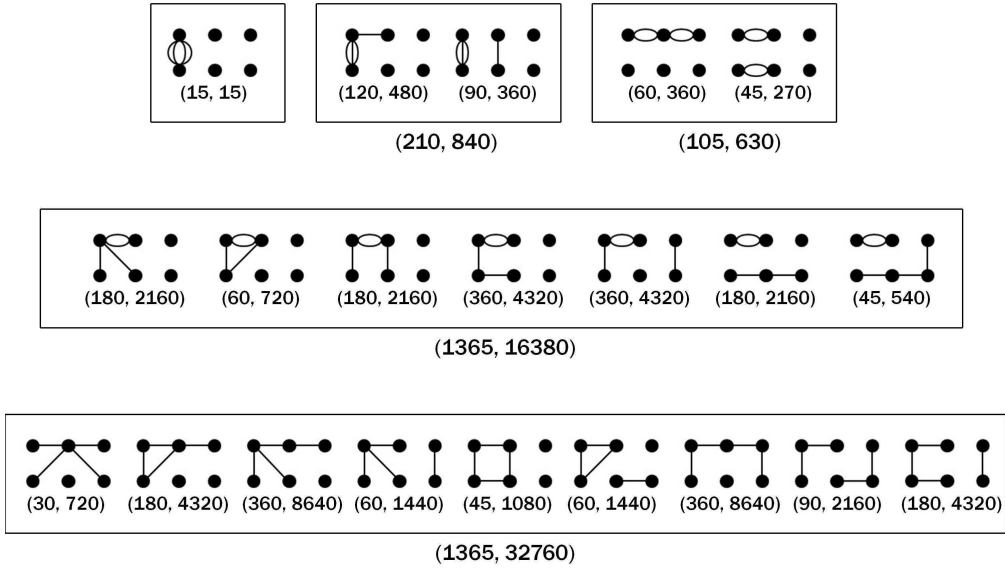


Figure 3: Number of labeled and fully labeled graphs for different isomorphisms and complexities with 6 vertices, 4 edges and no loops.

Table 5: Entropy and maximal entropy of graph data under three uniform random models for graphs with 6 vertices, 4 edges and no loops.

Data	Entropy of data according to			Maximal entropy
	ξ -model	η -model	ζ -model	
Fully labeled graph	15.63	15.45	14.67	15.63
Vertex labeled graph	11.44	11.58	11.07	11.58
Unlabeled graph	3.87	3.95	4.39	4.39
Graph Complexity	1.11	1.51	1.91	2.32

9 Other Graph Models

A natural generalization of the uniform model for the sequence $\boldsymbol{\xi} = (\xi_1, \dots, \xi_m)$ of sites for the edges is to assume that edges are independently assigned to sites according to a common arbitrary probability distribution

$$\mathbf{p} = (p_{ij} : (i, j) \in R)$$

over the possible sites of vertex pairs. Thus,

$$P(\boldsymbol{\xi} = \mathbf{g}) = \mathbf{p}^{\mathbf{m}(\mathbf{g})} = \prod_{(i,j) \in R} p_{ij}^{m_{ij}(\mathbf{g})} \quad \text{for } \mathbf{g} \in R^m .$$

The multiplicity sequence $\mathbf{m}(\boldsymbol{\xi})$ is multinomially distributed with parameters m and \mathbf{p} so that

$$P(\mathbf{m}(\boldsymbol{\xi}) = \mathbf{m}) = \binom{m}{\mathbf{m}} \mathbf{p}^{\mathbf{m}}$$

for the $\binom{m+r-1}{m}$ different ordered partitions \mathbf{m} of m into r non-negative integers. The complexity sequence $\mathbf{r}(\boldsymbol{\xi})$ has probabilities given by

$$P(\mathbf{r}(\boldsymbol{\xi}) = \mathbf{r}) = \sum_{\mathbf{m}|\mathbf{r}} \binom{m}{\mathbf{m}} \mathbf{p}^{\mathbf{m}} = \frac{m!}{\prod_{k=0}^m k!^{r_k}} \sum_{\mathbf{m}|\mathbf{r}} \mathbf{p}^{\mathbf{m}}$$

so, unless \mathbf{p} is uniform, the sum needs a specification of all multiplicity sequences that have complexity \mathbf{r} . It is straightforward to find the entropies:

$$H(\boldsymbol{\xi}) = -E \left[\log \mathbf{p}^{\mathbf{m}(\boldsymbol{\xi})} \right] = -m \mathbf{p} \log \mathbf{p} = m \sum_{(i,j) \in R} \varphi(p_{ij}) = m h(\mathbf{p})$$

and

$$\begin{aligned} H(\mathbf{m}(\boldsymbol{\xi})) &= -E \left[\log \binom{m}{\mathbf{m}(\boldsymbol{\xi})} \mathbf{p}^{\mathbf{m}(\boldsymbol{\xi})} \right] \\ &= m h(\mathbf{p}) - E \left[\log \binom{m}{\mathbf{m}(\boldsymbol{\xi})} \right] \\ &= m h(\mathbf{p}) - \log m! + \sum_{(i,j) \in R} E [\log m_{ij}(\boldsymbol{\xi})!] \\ &= m h(\mathbf{p}) - \log m! + \sum_{(i,j) \in R} \sum_{k=0}^m \binom{m}{k} p_{ij}^k (1 - p_{ij})^{m-k} \log k! . \end{aligned}$$

For $m > r$ there has to be some multiplicity larger than 1, but for $m \leq r$ it might be of interest to find the probability of no multiple edges. If loops are forbidden, this is the same as the probability of graph simplicity. If loops are allowed, the number of loops

$$m_1 = \sum_{i=1}^n m_{ii}$$

and the number of sites r_0 and r_1 with no and single occupancy are statistics that suffice to specify graph simplicity. For the ξ -model with common component distribution \mathbf{p} , the number of loops $m_1(\xi)$ is binomially distributed with parameters m and $p_1 = \sum_{i=1}^n p_{ii}$. The number $r_k(\xi)$ of sites with occupancy k has expected value

$$E[r_k(\xi)] = \sum_{(i,j) \in R} \binom{m}{k} p_{ij}^k (1 - p_{ij})^{m-k} \quad \text{for } k = 0, 1, \dots, m.$$

This implies the following expected values for the simple complexity measures given by the number of multiple occupancy sites and the number of multiple edges:

$$E[r - r_0(\xi) - r_1(\xi)] = r - \sum_{(i,j) \in R} (1 - p_{ij})^m - m \sum_{(i,j) \in R} p_{ij} (1 - p_{ij})^{m-1}$$

and

$$E[m - r_1(\xi)] = m \left(1 - \sum_{(i,j) \in R} p_{ij} (1 - p_{ij})^{m-1} \right).$$

The probability that there are no multiple edges is given by the sum of all ordered different products of m of the r probabilities in \mathbf{p} , that is by

$$P(r_1(\xi) = m) = m! \sum \prod_{(i,j) \in R} p_{ij}^{m_{ij}}$$

where the sum extends over all permutations of $\mathbf{m}^* = (0^{r-m} 1^m)$.

For many applications, an important generalization of independent assignments of edges to vertex pairs is obtained by introducing stochastic processes that generate edge sequences. A simple setup is to define r independent Poisson point processes that generate edges at the different sites with intensities λ_{ij} for $(i, j) \in R$. The sequence $\xi = (\xi_1, \dots, \xi_m)$ has components ξ_k that record the sites in the order the edges occur during a fixed period of time. Such an approach is to be discussed elsewhere.

An investigation of entropy measures for occupancy models similar to those considered here is described in an article by Frank and Nowicki (1989). They introduce a graph on objects corresponding to our edges with their edges specifying whether or not the objects

occupy the same site. Thus, this graph has complete connected components and is closely related to the concepts discussed here. They also develop asymptotic results for various entropies. Of special interest is the asymptotic entropy for the multinomial distribution, which implies that the multiplicities of the fully labeled graphs have an entropy $H(\mathbf{m}(\boldsymbol{\xi}))$ that for large m and r with r^2/m tending to zero is given by

$$H(\mathbf{m}(\boldsymbol{\xi})) = \frac{1}{2} \log \left[(2\pi em)^{r-1} \prod_{(i,j) \in R} p_{ij} \right] + O\left(\frac{r^2}{m}\right) .$$

Complexity is a general property considered in many different contexts and used with or without a specific definition. Complexity in graphs has been given different definitions in the literature focusing on other graph properties than edge multiplicity. For instance, Karreman (1955) and Mowshowitz (1968) are references that deal with completely different complexity properties of graphs used as models for molecules with chemical bonds between atoms. A common feature of many complexity concepts is that they seem to be well described and analyzed by information measures based on entropy.

References

- Carrington, P., Scott, J. and Wasserman, S. (eds.) (2005), *Models and Methods in Social Network Analysis*, New York: Cambridge University Press.
- Comtet, L. (1974), *Advanced Combinatorics: The Art of Finite and Infinite Expansions*, Dordrecht: Reidel Publishing Company.
- Frank, O. (2005), Network Sampling and Model Fitting, in *Models and Methods in Social Network Analysis*, eds. P. Carrington, J. Scott and S. Wasserman, New York: Cambridge University Press, 31–56.
- Frank, O. (2011a), Statistical Information Tools for Multivariate Discrete Data, in *Modern Mathematical Tools and Techniques in Capturing Complexity*, eds. L. Pardo, N. Balakrishnan and M. Ángeles Gil, Berlin: Springer Verlag, 177–190.
- Frank, O. (2011b), Survey Sampling in Networks, in *Handbook of Social Network Analysis*, eds J. Scott and P. Carrington, London: Sage Publications.
- Frank, O. and Nowicki, K. (1989), On Entropies of Occupancy Distributions, in *Combinatorics and Graph Theory*, eds. Z. Skupien, M. Borowiecki, Warsaw: Banach Center Publications, **25**, PWN-Polish Scientific Publishers, 71–86.
- Karreman, G. (1955), Topological Information Content and Chemical Reactions, *Bulletin of Mathematical Biophysics*, **17**, 279–285.
- Kolaczyk, E. (2009), *Statistical Analysis of Network Data*, New York: Springer Verlag.
- Meyers, R. (ed.) (2009), *Encyclopedia of Complexity and Systems Science*, New York: Springer Verlag.
- Mowshowitz, A. (1968), Entropy and the Complexity of Graphs: I. An Index of the Relative Complexity of a Graph, *Bulletin of Mathematical Biophysics*, **30**, 175–204.
- Scott, J. and Carrington, P. (eds.) (2011), *Handbook of Social Network Analysis*, London: Sage Publications.

Random Stub Matching Models of Multigraphs

Termeh Shafie

Abstract

This article studies the local and global structure of multigraphs under random stub matching with fixed degrees (RSM). The local structure is analyzed by marginal distributions of edge multiplicities, and the global structure is analyzed by the simultaneous distribution of edge multiplicities. The simultaneous distribution is shown to depend on a single complexity statistic. The distributions under RSM and IEA are used for calculations of moments and entropies, and for comparisons by information divergence. The modified distributions are obtained by ignoring the dependencies between edges and assuming independent edge assignments to sites (IEA), and by ignoring the dependencies between stubs and assuming independent stub assignments to vertices (ISA). The main results in this article include a new formula for the probability of an arbitrary number of loops at a vertex, and a more intricate expression for the probability of an arbitrary number of edges at any site. Further, simplicity and complexity of multigraphs under RSM are investigated and a new method of approximating the probability that an RSM multigraph is simple is proposed and shown to perform well for multigraphs with small numbers of vertices and edges.

Keywords: multigraph, edge multiplicity, entropy, information divergence, simplicity and complexity.

1 Introduction

It is well known that different degree sequences are compatible with different numbers of graphs. Several methods have been developed for generating random graphs with fixed or modeled degrees, degree distributions or expected degrees. Such methods can be found in Blitzstein and Diaconis (2011), Bayati, Kim and Saberi (2010), Britton, Deijfen and Martin-Löf (2006), Chung and Lu (2002), and Bender and Canfield (1978). Random stub matching, also referred to as the configuration model or the pairing model (e.g. Janson 2009, Bollobàs 1980), generates random multigraphs by randomly coupling pairs of stubs to form edges.

Department of Statistics, Stockholm University, S-106 91 Stockholm, termeh.shafie@stat.su.se

This article focuses on both the local and global structure of multigraphs under random stub matching with fixed degrees (RSM). The local structure is analyzed by marginal distributions of edge multiplicities, and the global structure is analyzed by the simultaneous distribution of edge multiplicities. The distributions under RSM as well as some modified distributions with modeled degrees are used for calculations of moments and entropies, and for comparisons by information divergences. These modified distributions are obtained by ignoring the dependencies between edges when they are assigned to sites (IEA) and by ignoring the dependencies between stubs when they are assigned to vertices (ISA).

In the next section, some basic concepts such as stubs, edges, sites, multigraphs, multiplicities and complexities are presented. The uniform random stub matching procedure given fixed degrees is described in Section 3 where the distribution of multigraphs is determined and shown to depend on a single statistic which is a special summary measure of complexity.

The moments of the edge multiplicity distributions under RSM are derived in Section 4. The moments of the number of loops at a fixed vertex are determined as functions of the number of edges, denoted m , and the degree at that vertex. It is shown that the variance of the number of loops under RSM is less than the variance under IEA, except for the degenerate cases of degree value 1 or $2m$. The moments of the number of edges between two distinct vertices are determined as functions of the total number of edges m and the degrees at the two vertices. It is shown that the variance of the number of such edges under RSM is generally less than the variance under IEA, except for degrees that lie symmetrically around the total number of edges and are given by $m \pm k$ for any non-negative integer k less than a specified limit.

In Section 5, the distributions of edge multiplicities at local sites are investigated. The probability of no loops at a vertex has been given in the literature (Janson 2009) but, as far as we know, not generalized to arbitrary number of loops at a vertex. This formula is derived using a special technique which also allows us to find the probability of edge frequency between pairs of distinct vertices. This technique gives the trivariate distribution of the numbers of loops and non-loops within and between two distinct vertices, and from its marginals we obtain distributions of local edge multiplicities at any site. Although somewhat combinatorically tedious, we also determine the range space of the trivariate distribution.

Throughout Section 5, information divergence and entropies are used to compare the edge multiplicity distributions under RSM and IEA. Numerical examples using the divergence indicate that the distribution of loop multiplicity under RSM is closely related to that of the IEA distribution at vertices with low degrees. The divergence increases monotonically from zero to a maximal value and decreases very steeply back to zero. The results also indicate that the discrepancy between the edge multiplicity distributions under RSM and IEA is due to their different range spaces. Further, an illustration is given of how the divergence between the probability distributions of local edge frequency under RSM and

IEA varies for different degrees. The results also show how the resemblance between the distributions increases with increasing m .

The flatness of the local edge multiplicity distributions under RSM and IEA are compared using entropies. Specifically, the entropy of the loop multiplicity distribution under RSM is shown to be more symmetrical than that of IEA, and it has its maximum around the stub proportion value 0.5. The corresponding loop multiplicity distribution under IEA is skew to the right and has its maximum for a stub proportion value of about 0.7. Good approximations to the entropies of loop multiplicities under both RSM and IEA are found.

Special attention is paid to the edge multiplicity distribution for the case with two degrees that lie symmetrically around m . For this case the entropies are much higher for IEA than for RSM. For both RSM and IEA, we give approximations to the entropies. These approximations are very good for the IEA distributions but not for the RSM distributions.

In Section 6 the global structure is analyzed by the distribution of multigraphs under RSM and IEA. Under RSM, this distribution was earlier shown to depend on a single complexity statistic, and in order to find the entropy of this distribution, results about edge multiplicities from Section 5 are used. The approximate entropies of the RSM and IEA distribution of multigraphs are given using covariance matrices. For both RSM and IEA, the exact and approximate entropies are close to the upper bounds of the exact entropies. Using the information divergence, a large deviation between the multigraph distributions under RSM and IEA are found and the results indicate flat distributions over very different ranges. In particular for regular multigraphs, both the RSM and IEA distribution cluster at the high probability sites when more edges are added and are therefore less flat for large values of m .

In the final section, the simplicity and complexity of multigraphs under RSM are studied. Two asymptotic results for the probability that an RSM multigraph is simple are numerically investigated, and it is shown that these probability approximations do not perform well for multigraphs with small numbers of vertices and edges. Under certain conditions, an alternative way of approximating the probability that an RSM multigraph is simple is proposed. Numerical examples show that this approximation is good for small multigraphs. Some other variables that identify simplicity and complexity are also considered, and the moments of these variables are derived. It is shown that the moments of some of these variables are much easier handled under IEA and a convenient way of obtaining the IEA distribution is introduced. This is done by assuming that the stubs are randomly generated and independently assigned to sites (ISA) and can be viewed as a Bayesian model for the stub frequencies under RSM. Using this method, approximations to the entropy of the distribution of multigraphs under RSM are derived. If the degree distributions are uniformly or close to uniformly distributed, the approximations are good even for small multigraphs, and for skew distributions they are good for multigraphs with many edges. An asymptotic equipartition property is shown to give alternative approximations that work reasonably well except for multigraphs with skew degree sequences and few vertices.

2 Basic Concepts and Notation

A finite undirected graph g with n labeled vertices and m labeled edges associates with each edge an ordered or unordered vertex pair. Let $V = \{1, \dots, n\}$ and $E = \{1, \dots, m\}$ be the sets of vertices and edges labeled by integers, and denote by R the set of available sites for the edges. The site space for directed edges is V^2 and the site space for undirected edges is $R = \{(i, j) \in V^2 : i \leq j\}$. We consider (i, j) with $i \leq j$ as a canonical representation for the unordered vertex pair. Let $r = \binom{n+1}{2}$ be the number of sites in R .

The degree of vertex i , the number of edges incident to it, is denoted d_i and $\sum_i^n d_i = 2m$. The degree sequence $\mathbf{d} = (d_1, \dots, d_n)$ defines another sequence of $2m$ vertices or edge-stubs corresponding to m edges without specifying the pairings of stubs to edges:

$$\mathbf{s} = (\underbrace{1 \dots 1}_{d_1} \quad \underbrace{2 \dots 2}_{d_2} \quad \dots \quad \underbrace{n \dots n}_{d_n}) .$$

Thus there is a bijection $\mathbf{d} \leftrightarrow \mathbf{s}$ and we use the shorthand notation

$$\mathbf{s} = (s_1, \dots, s_{2m}) = (1^{d_1} 2^{d_2} \dots n^{d_n}) \in V^{2m} .$$

Let $X(\mathbf{d})$ be the set of sequences \mathbf{x} that are permutations of the stub sequence

$$X(\mathbf{d}) = \{\mathbf{x} = (x_1, \dots, x_{2m}) \in V^{2m} : \mathbf{x} \sim \mathbf{s}\} ,$$

where \sim means "is a permutation of". The number of permutations of a stub sequence \mathbf{s} obtained from the degree sequence \mathbf{d} is given by

$$|X(\mathbf{d})| = \binom{2m}{\mathbf{d}} = \frac{(2m)!}{\mathbf{d}!} = \frac{(2m)!}{\prod_{i=1}^n d_i!} ,$$

and is also denoted $\#(\mathbf{x}|\mathbf{s})$ or $\#(\mathbf{x}|\mathbf{d})$.

Edges at the same site are called multiple edges, and the number of multiple edges at site (i, j) is its multiplicity denoted m_{ij} . The multiplicity at site $(i, j) \in V^2$ in \mathbf{x} is

$$m_{ij}(\mathbf{x}) = \sum_{k=1}^m I((x_{2k-1}, x_{2k}) = (i, j)) .$$

It follows that

$$\sum_{i=1}^n \sum_{j=1}^n m_{ij}(\mathbf{x}) = m_{..}(\mathbf{x}) = m$$

and

$$\sum_{j=1}^n (m_{ij}(\mathbf{x}) + m_{ji}(\mathbf{x})) = m_{i.}(\mathbf{x}) + m_{.i}(\mathbf{x}) = d_i \quad \text{for } i = 1, \dots, n .$$

The number of loops and the number of non-loops are denoted

$$m_1(\mathbf{x}) = \sum_{i=1}^n m_{ii}(\mathbf{x}) \quad \text{and} \quad m_2(\mathbf{x}) = \sum_{i \neq j} m_{ij}(\mathbf{x}).$$

When the edge multiplicities in \mathbf{x} are arranged as a matrix we obtain the edge multiplicity matrix

$$\mathbf{m}(\mathbf{x}) = (m_{ij}(\mathbf{x}) : (i, j) \in V^2)$$

with loop counts $m_{ii}(\mathbf{x})$ in the main diagonal. If a matrix is created with these loop counts as the elements in the main diagonal and zeros outside, we obtain the loop frequency matrix $\mathbf{m}_1(\mathbf{x})$. The non-loop frequency matrix denoted $\mathbf{m}_2(\mathbf{x})$ is then given as $\mathbf{m}_2(\mathbf{x}) = \mathbf{m}(\mathbf{x}) - \mathbf{m}_1(\mathbf{x})$.

The representation of the edge sequence is modified in two different ways. The sequence $\mathbf{y} = (y_1, \dots, y_{2m})$ is obtained from \mathbf{x} by vertex shifts according to $(y_{2k-1}, y_{2k}) = (\min(x_{2k-1}, x_{2k}), \max(x_{2k-1}, x_{2k}))$ for $k = 1, \dots, m$. In other words, the injective map $\mathbf{x} \rightarrow \mathbf{y}$ gives an ordered sequence of m edges from R and \mathbf{y} is the edge sequence of an undirected graph generated from \mathbf{x} . The number of \mathbf{x} that yield the same \mathbf{y} is then given as $\#(\mathbf{x}|\mathbf{y}) = 2^{m_2(\mathbf{y})}$. The sequence $\mathbf{z} = (z_1, \dots, z_{2m})$ is obtained from \mathbf{y} by ordering its edges non-decreasingly, i.e. the injective map $\mathbf{y} \rightarrow \mathbf{z}$ gives an edge sequence canonically ordered according to

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

so that

$$(z_1, z_2) \leq (z_3, z_4) \leq \dots \leq (z_{2m-1}, z_{2m}).$$

The edge sequence \mathbf{z} represents the vertex labeled graph given by \mathbf{y} without the edge labels. The set of sequences \mathbf{z} generated by $\mathbf{x} \in X(\mathbf{d})$ is denoted $Z(\mathbf{d})$. The number of \mathbf{x} that yield the same vertex labeled graph \mathbf{z} is $\#(\mathbf{x}|\mathbf{z}) = \sum_{\mathbf{y}|\mathbf{z}} \#(\mathbf{x}|\mathbf{y}) = \sum_{\mathbf{y}|\mathbf{z}} 2^{m_2(\mathbf{y})}$. Now $m_2(\mathbf{y}) = m_2(\mathbf{z})$ and

$$\#(\mathbf{y}|\mathbf{z}) = \binom{m}{\mathbf{m}(\mathbf{z})} = \frac{m!}{\prod_{i \leq j} m_{ij}(\mathbf{z})!},$$

so that

$$\#(\mathbf{x}|\mathbf{z}) = 2^{m_2(\mathbf{z})} \binom{m}{\mathbf{m}(\mathbf{z})}.$$

Here the edge multiplicity at site $(i, j) \in V^2$ in \mathbf{y} and \mathbf{z} is equal to the common value

$$m_{ij}(\mathbf{y}) = m_{ij}(\mathbf{z}) = \begin{cases} m_{ij}(\mathbf{x}) & \text{for } i = j \\ m_{ij}(\mathbf{x}) + m_{ji}(\mathbf{x}) & \text{for } i < j \\ 0 & \text{for } i > j. \end{cases}$$

The common edge multiplicity matrices of \mathbf{y} and \mathbf{z} are triangular with zeros below the main diagonal. Moreover, the loop frequency matrices of \mathbf{x} , \mathbf{y} and \mathbf{z} are all equal. The numbers of loops and non-loops are therefore equal to

$$m_1(\mathbf{z}) = \sum_{i=1}^n m_{ii}(\mathbf{z}) = \sum_{i=1}^n m_{ii}(\mathbf{x})$$

and

$$m_2(\mathbf{z}) = \sum_{i<j} \sum m_{ij}(\mathbf{z}) = \sum_{i \neq j} \sum m_{ij}(\mathbf{x}) .$$

The sum of the multiplicity matrix and its transpose is the same symmetric matrix for \mathbf{x} , \mathbf{y} and \mathbf{z} , i.e.

$$\mathbf{m}(\mathbf{x}) + \mathbf{m}'(\mathbf{x}) = \mathbf{m}(\mathbf{y}) + \mathbf{m}'(\mathbf{y}) = \mathbf{m}(\mathbf{z}) + \mathbf{m}'(\mathbf{z}) .$$

The row and column sums in this matrix are given by the degrees, and the loop counts are doubled in the main diagonal.

Figure 1 shows a schematic view of bijections and other functional relationships between the various concepts introduced here. The functional relationships comprise three different edge sequences and their edge multiplicity matrices, the stub sequence, and the degree sequence.

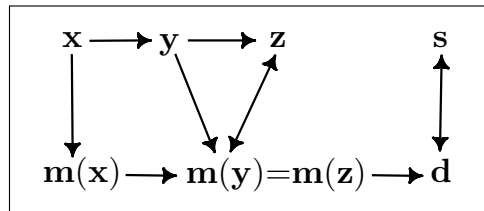


Figure 1: Functional relationships between the edge sequences, multiplicity matrices, stub sequence and degree sequence.

3 Uniform Stub Matching with Fixed Degrees

We focus on uniform distributions for different families of graphs which we refer to as random graphs. Assume that ξ is a random permutation of the stub sequence \mathbf{s} defined by the degree sequence \mathbf{d} , i.e. ξ is uniform on $X(\mathbf{d}) = \{\mathbf{x} : \mathbf{x} \sim \mathbf{s}\} \subseteq V^{2m}$ with probabilities

$$P(\xi = \mathbf{x}) = \frac{1}{\binom{2m}{\mathbf{d}}} \quad \text{for } \mathbf{x} \in X(\mathbf{d}) .$$

Let $\boldsymbol{\eta}$ be the edge sequence of the undirected graph obtained by shifts in $\boldsymbol{\xi}$. Further let ζ be the canonical edge sequence of the undirected graph generated by $\boldsymbol{\xi}$ with probabilities

$$\begin{aligned} P(\zeta = \mathbf{z}) &= \sum_{\mathbf{x}|\mathbf{z}} P(\boldsymbol{\xi} = \mathbf{x}) = \frac{2^{m_2(\mathbf{z})} \binom{m}{\mathbf{m}(\mathbf{z})}}{\binom{2m}{\mathbf{d}}} = \frac{2^{m_2(\mathbf{z})} m! \mathbf{d}!}{\mathbf{m}(\mathbf{z})! (2m)!} \\ &= \frac{2^{m_2(\mathbf{z})} m! \prod_{i=1}^n d_i!}{(2m)! \prod_{i \leq j} m_{ij}(\mathbf{z})!} \quad \text{for } \mathbf{z} \in Z(\mathbf{d}) . \end{aligned}$$

Consider the ordered partition $\mathbf{m}(\mathbf{z})$ and the corresponding unordered partition of m into r non-negative integers. There is a bijection between this partition and the sequence of frequencies of sites with multiplicities $0, 1, \dots, m$ given by $\mathbf{r}(\mathbf{z}) = (r_0(\mathbf{z}), \dots, r_m(\mathbf{z}))$ where

$$r_k(\mathbf{z}) = \sum_{i \leq j} I(m_{ij}(\mathbf{z}) = k) \quad \text{for } k = 0, 1, \dots, m .$$

The distribution of multiplicities that is given by $\mathbf{r}(\mathbf{z})$ is called the complexity of the graph with edge sequence \mathbf{z} (Frank and Shafie, 2012). It is convenient to separate frequencies of loops and non-loops and use $r(\mathbf{z}) = r_1(\mathbf{z}) + r_2(\mathbf{z})$ where

$$\mathbf{r}_1(\mathbf{z}) = (r_{10}(\mathbf{z}), \dots, r_{1m}(\mathbf{z})) \quad \text{and} \quad \mathbf{r}_2(\mathbf{z}) = (r_{20}(\mathbf{z}), \dots, r_{2m}(\mathbf{z}))$$

with

$$r_{1k}(\mathbf{z}) = \sum_{i=1}^n I(m_{ii}(\mathbf{z}) = k) \quad \text{and} \quad r_{2k}(\mathbf{z}) = \sum_{i < j} I(m_{ij}(\mathbf{z}) = k) \quad \text{for } k = 0, 1, \dots, m .$$

Using these complexities it is possible to express the probability $P(\zeta = \mathbf{z})$ as a function of a special summary measure of complexity according to the following:

$$P(\zeta = \mathbf{z}) = C 2^{-t(\mathbf{z})} ,$$

where $C = 2^m m! \mathbf{d}! / (2m)!$ and

$$\begin{aligned} t(\mathbf{z}) &= m_1(\mathbf{z}) + \log \mathbf{m}(\mathbf{z})! \\ &= \sum_{i=1}^n m_{ii}(\mathbf{z}) + \sum_{i \leq j} \log m_{ij}(\mathbf{z})! \\ &= \sum_{k=1}^m k r_{1k}(\mathbf{z}) + \sum_{k=2}^m r_k(\mathbf{z}) \log k! \\ &= \sum_{k=1}^m (k + \log k!) r_{1k}(\mathbf{z}) + \sum_{k=2}^m r_{2k}(\mathbf{z}) \log k! . \end{aligned}$$

Simple graphs without loops and multiple edges have $t(\mathbf{z}) = 0$ and all simple graphs have the same probability C . More complex graphs have higher values of $t(\mathbf{z})$ and smaller probabilities. All graphs with a fixed value $t(\mathbf{z}) = t$ of the complexity measure have the same probability. The set $Z(\mathbf{d})$ of edge sequences is partitioned according to values of the complexity measure, and the set of edge sequences with complexity t is denoted

$$Z(\mathbf{d}, t) = \{\mathbf{z} \in Z(\mathbf{d}) : t(\mathbf{z}) = t\} .$$

The number of sequences in this set is denoted $|Z(\mathbf{d}, t)| = K(\mathbf{d}, t)$, or simply K_t if \mathbf{d} is clear from context. The probability of complexity value t is given by

$$P(t(\zeta) = t) = \sum_{\mathbf{z}|t(\mathbf{z})=t} P(\zeta = \mathbf{z}) = CK_t 2^{-t} ,$$

and the conditional distribution of ζ given complexity t is equal to

$$P(\zeta = \mathbf{z} | t(\zeta) = t) = \frac{C 2^{-t}}{CK_t 2^{-t}} = \frac{1}{K_t}$$

which is uniform on K_t outcomes in $Z(\mathbf{d}, t)$. Neither K_t nor the probability $P(t(\zeta) = t)$ are monotone as functions of t . This will follow by an examination of K_t in a numerical example considered below.

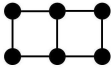
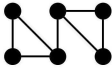
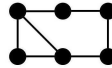
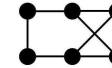
Example.

Consider undirected graphs with degree sequence $\mathbf{d} = (3, 3, 2, 2, 2, 2)$. There are in total 784 possible vertex labeled graphs with edge sequences \mathbf{z} in $Z(\mathbf{d})$. Table 1 lists the probability $P(t(\zeta) = t)$, the number of vertex labeled graphs K_t , and the probability per graph $P(t(\zeta) = t)/K_t$ for each complexity value 2^t . As seen, neither K_t nor the probability $P(t(\zeta) = t)$ are monotone as functions of t . It is also clear that every simple graph has a higher probability of occurring than any complex graph, and that the more complex a graph is, the smaller probability it has. Note however that this does not mean that all unlabeled graphs of given complexity have the same probability of occurring. This is clarified by looking at all unlabeled simple graphs in this example. In Table 2 the number of isomorphic graphs (the number of vertex labeled graphs) are listed for the four possible unlabeled simple graphs. We see in Table 2 that the third unlabeled graph has more edge sequences generating it than the others, and therefore it has the highest probability of occurring.

Table 1: Complexity distribution of graphs with degree sequence (3, 3, 2, 2, 2, 2).

Complexity (2^t)	Probability of complexity	Number of graphs	Probability per graph
1	0.230170	54	0.004262
2	0.387880	182	0.002131
4	0.252522	237	0.001065
6	0.002131	3	0.000710
8	0.098035	184	0.000533
12	0.001421	4	0.000355
16	0.024242	91	0.000266
24	0.000535	3	0.000178
32	0.002398	18	0.000133
48	0.000533	6	0.000089
64	0.000067	1	0.000067
96	0.000044	1	0.000044

Table 2: Simple graphs with degree sequence (3, 3, 2, 2, 2, 2).

Unlabeled graphs					Total
Vertex labeled graphs	12	6	24	12	54

4 Moments of Edge Multiplicities

In order to investigate the distribution of the edge multiplicities under random stub matching (RSM), we start by analyzing the moments of this distribution. The probability of coupling stubs to edges in ξ is

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

where $\sum_{i=1}^n \sum_{j=1}^n P_{ij} = 1$. The probability of undirected edges in $\boldsymbol{\eta}$ is thus equal to

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

Note that (η_{2k-1}, η_{2k}) are identically but not independently distributed. It is convenient to introduce $Q_{ijkl} = P((\eta_{2u-1}, \eta_{2u}) = (i, j) \text{ and } (\eta_{2v-1}, \eta_{2v}) = (k, \ell))$ for $u \neq v$. For $i \leq j$ and $k \leq \ell$ we have that $Q_{ijkl} = Q_{klij}$.

The expected values and variances of the numbers of loops and non-loops in $\boldsymbol{\eta}$ (or in the canonical edge sequence $\boldsymbol{\zeta}$) under RSM are derived by using the first and second moments of the edge multiplicities, which occasionally is shortly denoted m_{ij} when randomness is clear, i.e.

$$m_{ij} = m_{ij}(\boldsymbol{\eta}) = m_{ij}(\boldsymbol{\zeta}) = \sum_{k=1}^m I_{ijk},$$

where the indicators are given by

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

for $(i, j) \in R$ and $k \in E$. Now $E(I_{ijk}) = Q_{ij}$ so that

$$E(m_{ij}) = mQ_{ij} = \begin{cases} \binom{d_i}{2} / (2m - 1) & \text{for } i = j \\ d_i d_j / (2m - 1) & \text{for } i < j , \end{cases}$$

In order to obtain the variance of m_{ij} under RSM, we need the covariance between indicators I_{ijk} and $I_{ij\ell}$. They are given by

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

where

$$Q_{ijij} = \begin{cases} \binom{d_i}{2} \binom{d_i-2}{2} / \binom{2m}{2} \binom{2m-2}{2} & \text{for } i = j \\ d_i d_j (d_i - 1)(d_j - 1) / \binom{2m}{2} \binom{2m-2}{2} & \text{for } i < j . \end{cases}$$

Hence

$$\begin{aligned}
\text{Var}(m_{ij}) &= \sum_{k=1}^m \sum_{\ell=1}^m \text{Cov}(I_{ijk}, I_{ij\ell}) \\
&= mQ_{ij}(1 - Q_{ij}) + m(m-1)(Q_{ijij} - Q_{ij}^2) \\
&= mQ_{ij}(1 - mQ_{ij}) + m(m-1)Q_{ijij} \\
&= \begin{cases} \frac{\binom{d_i}{2}}{2m-1} \left(1 - \frac{\binom{d_i}{2}}{2m-1}\right) + \frac{6\binom{d_i}{4}}{(2m-1)(2m-3)} & \text{for } i = j \\ \frac{d_i d_j}{(2m-1)} \left(1 - \frac{d_i d_j}{2m-1}\right) + \frac{d_i d_j (d_i-1)(d_j-1)}{(2m-1)(2m-3)} & \text{for } i < j . \end{cases}
\end{aligned}$$

Covariances between m_{ij} and $m_{k\ell}$ require covariances between indicators I_{iju} and $I_{k\ell v}$ for $u = 1, \dots, m$ and $v = 1, \dots, m$. Here $i \leq j$ and $k \leq \ell$ and, since $\text{Cov}(m_{ij}, m_{k\ell}) = \text{Cov}(m_{k\ell}, m_{ij})$, it is sufficient to consider $i \leq k$, and for $i = k$, only $j \leq \ell$. Explicit expressions for such covariances will be given when needed in the sequel.

The variance of m_{ij} under RSM can be written as

$$\text{Var}(m_{ij}) = \sigma_{ij}^2 + \Delta_{ij} \quad \text{for } i \leq j .$$

Here, $\sigma_{ij}^2 = mQ_{ij}(1 - Q_{ij})$ is the variance of a binomial distribution obtained by independent edge assignments (IEA) with parameters m and Q_{ij} for $i \leq j$ and

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

Using these expressions we can now 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 of edge multiplicity. We start with the case when $i = j$ and search the sign of Δ_{ii} for values of $d_i = d$ where $2 \leq d \leq 2m - 1$. By rewriting Δ_{ii} as

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

and noticing that

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

is a decreasing function of k , it follows that $\Delta_{ii} < 0$ for $d < 2m$. Thus, $\text{Var}(m_{ii}) < \sigma_{ii}^2$ for $1 < d < 2m$ and $\text{Var}(m_{ii}) = \sigma_{ii}^2$ only for the degenerate cases $d = 1$ and $d = 2m$ with $\sigma_{ii}^2 = 0$.

When $i < j$, set $a = \min(d_i, d_j)$ and $b = \max(d_i, d_j)$ and search the sign of Δ_{ij} for different pairs of values (a, b) with $1 \leq a \leq b$ and $a + b \leq 2m$ for $m > 1$. By rewriting Δ_{ij} as

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

we see that Δ_{ij} has the same sign as the function

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

Now $f(a, b) < 0$ for $1 \leq a \leq b \leq m-1$, and $f(1, b) < 0$ for $1 \leq b \leq 2m-1$. 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

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

where $\theta = (4m-3)/m(2m-1)$ and between 0 and 1. The intersection between this curve and the upper boundary $b = 2m-a$ of the (a, b) -region defined by $1 \leq a \leq b$ and $a+b \leq 2m$ 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

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

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

and

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

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 $1 \leq a \leq b \leq 2m-1-a$. The conclusion is that

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

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. Thus the RSM distribution of m_{ij} has a variance that is smaller than σ_{ij}^2 unless d_i and d_j lie symmetrically around m and are given by $m \pm k$ for some non-negative integer

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

It also follows that the variance is maximal for $k = 0$ and decreases for increasing k . The case $m = 20$ is illustrated in Figure 2 where the points with positive Δ_{ij} are marked with (*).

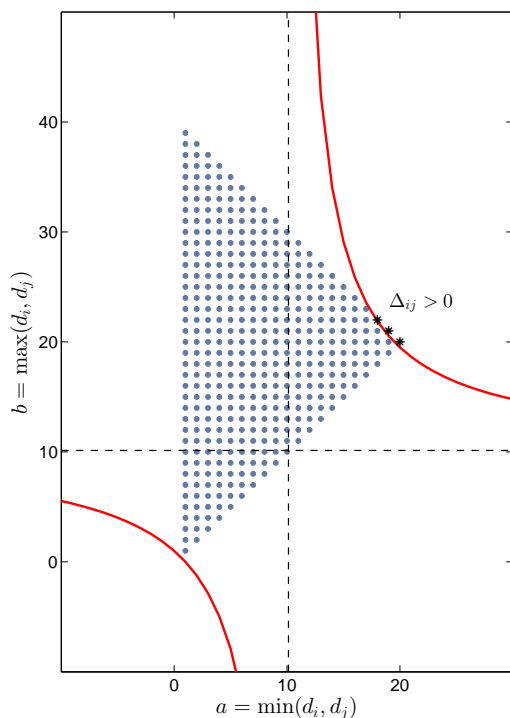


Figure 2: Points and critical curve. Points represent possible degree pairs (a, b) at a given vertex pair in a graph with $m = 20$ edges. A critical curve is separating points with positive and negative variance difference Δ_{ij} between the edge multiplicity distributions obtained at (a, b) by random stub matching and by independent edge assignments. The points where the stub matching has larger variance are marked by (*).

5 Distributions of Edge Multiplicities

In this section, some new results on distributions of edge multiplicities for graphs obtained by random stub matching (RSM) are derived. There are many asymptotic results in the literature (e.g. Bollobàs 2001) but we are also interested in exact results for fixed degree sequences. Specific results for no loops and no multiple edges have been discussed in Janson (2009) and Bollobàs (1980), where random stub matching is referred to as the configuration model or the pairing model.

The probability of no loops at vertex i , denoted P_0 , is given by Janson (2009) as

$$P_0 = \prod_{j=1}^{d_i} \frac{2m - d_i - j + 1}{2m - 2j + 1} .$$

He gives no formula for arbitrary numbers of loops at i but notes that it is more difficult to find the probability of no multiple edges due to the complications caused by loops. We now derive a formula for an arbitrary number of loops at vertex i under RSM and get in particular a simple expression for the number of no loops at i . This technique can be generalized and used to derive the probability of arbitrary multiplicities at any site $(i, j) \in R$.

Consider the probability of v loops at vertex i under RSM denoted by $P_v = P(m_{ii}(\boldsymbol{\eta}) = v) = P(m_{ii} = v)$ for $v = 0, \dots, m$. To find how many of the $\binom{2m}{\mathbf{d}}$ possible stub sequences that generate v loops at i , arrange m edges with v loops at i , $d_i - 2v$ edges with the remaining i -stubs, and $m - d_i + v$ other edges. This number of arrangements is given by the multinomial coefficient $\binom{m}{v, d_i - 2v}$. The single i -stubs have two alternative locations in the $d_i - 2v$ 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

$$P_v = \frac{\binom{m}{v, d_i - 2v} 2^{d_i - 2v} \binom{2m - d_i}{\mathbf{d}^*}}{\binom{2m}{\mathbf{d}}}$$

which simplifies to

$$P_v = \frac{\binom{m}{v, d_i - 2v} 2^{d_i - 2v}}{\binom{2m}{d_i}} .$$

In particular the probability of no loops at vertex i under RSM is equal to

$$P_0 = \frac{\binom{m}{d_i} 2^{d_i}}{\binom{2m}{d_i}} .$$

This formula can be developed according to the following which shows that it is equivalent to Janson's (2009) expression for P_0 as a ratio between a falling factorial from $2m - d_i$ and

a falling semifactorial from $2m - 1$, both carried out for d_i factors (in fact, $d_i - 1$ factors suffice since the last one cancels):

$$\begin{aligned}
P_0 &= \frac{m! d_i! (2m - d_i)! 2^{d_i}}{d_i! (m - d_i)! (2m)!} \\
&= \frac{m! 2^m (2m - d_i)!}{(2m)! (m - d_i)! 2^{m-d_i}} \\
&= \frac{(2m)!! (2m - d_i)!}{(2m)! (2m - 2d_i)!!} \\
&= \frac{(2m - 2d_i - 1)!! (2m - d_i)!}{(2m - 1)!! (2m - 2d_i)!} \\
&= \frac{(2m - d_i)(2m - d_i - 1) \cdots (2m - 2d_i + 1)}{(2m - 1)(2m - 3) \cdots (2m - 2d_i + 1)} .
\end{aligned}$$

Assume that $d_i = d$ with $2 \leq d \leq 2m$ and consider the general probability that there are v loops at vertex i under RSM given by

$$P(m_{ii} = v) = \frac{\binom{m}{v, d-2v} 2^{d-2v}}{\binom{2m}{d}} \quad \text{for } v = 0, 1, \dots, \lfloor d/2 \rfloor .$$

This probability is also denoted P_v or $P_v(m, d)$. Set $\mathbf{P} = (P_0, \dots, P_{\lfloor d/2 \rfloor})$. The expected value and variance of m_{ii} under RSM given in the previous section are equal to μ_{ii} and $\sigma_{ii}^2 + \Delta_{ii}$, where $\mu_{ii} = \binom{d}{2}/(2m - 1)$ and $\sigma_{ii}^2 = mQ_{ii}(1 - Q_{ii}) = \mu_{ii} \left(1 - \frac{\mu_{ii}}{m}\right)$ are the mean and variance of the IEA distribution $\mathbf{B} = (B_0, \dots, B_m)$ with parameters m and Q_{ii} . The range of the multiplicity distribution under IEA is $v = 0, 1, \dots, m$ and the range of the multiplicity distribution under RSM is smaller. Its proportion is $(\lfloor d/2 \rfloor + 1)/(m + 1)$ of the range of the IEA distribution. Table 3 gives these distributions for the case $m = 10$ and $d = 10$. Also presented in Table 3 is a measure of the discrepancy between the distributions given by the information divergence

$$D(\mathbf{P}, \mathbf{B}) = \sum_{v: P_v > 0} P_v \log \frac{P_v}{B_v} .$$

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 (see e.g. Frank 2011). Figure 3 shows how divergence varies for different values of $d = 2, \dots, 2m - 1$ for $m = 40$, and how the divergence varies for different stub proportions $d/2m$ (or range proportions $(\lfloor d/2 \rfloor + 1)/(m + 1)$) for some values of m .

Table 3: The probability distribution of loop multiplicity at a vertex of degree $d = 10$ when $m = 10$ edges are formed by random stub matching (RSM). It is compared by information divergence to a binomial distribution obtained by $m = 10$ independent edge assignments (IEA).

Number of loops	Probability under RSM	Probability under IEA	Weighted log-likelihood ratio
0	0.005542	0.067011	-0.019930
1	0.124705	0.207964	-0.092044
2	0.436468	0.290433	0.256636
3	0.363723	0.240358	0.217242
4	0.068198	0.130539	-0.063849
5	0.001364	0.048615	-0.007164
6	0	0.012573	0
7	0	0.002230	0
8	0	0.000259	0
9	0	0.000018	0
10	0	0.000001	0

Divergence = 0.290891

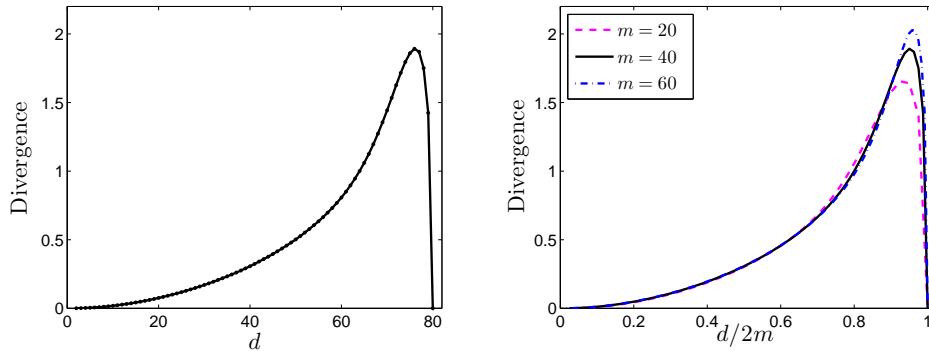


Figure 3: Information divergence between random stub matching and independent edge assignments for the distributions of loop multiplicity at a vertex of degree d in a graph with m edges. Information divergence is plotted against degree d for $m = 40$ and against stub proportion $d/2m$ for $m = 20, 40, 60$.

In order to compare the two distributions we also use the entropy which characterize flatness of the distributions \mathbf{P} and \mathbf{B} . The entropies are equal to

$$h(\mathbf{P}) = \sum_{v:P_v>0} -P_v \log P_v$$

and

$$h(\mathbf{B}) = \sum_{v:B_v>0} -B_v \log B_v .$$

Upper bounds to the entropies are given by their logarithmic ranges:

$$h(\mathbf{P}) \leq \log (\lfloor d/2 \rfloor + 1)$$

and

$$h(\mathbf{B}) \leq \log(m + 1) .$$

For the case where $m = 10$ and $d = 10$ we have that $h(\mathbf{P}) = 1.746$ and $h(\mathbf{B}) = 2.443$. Figure 4 shows how entropies vary for different stub proportions $d/2m$ for $m = 20, 40, 60$. We see that stub matching has lower entropy and is more symmetric around 0.5 than the entropy corresponding to independent edge assignments. The latter entropy is skew to the right and has its maximum when loop probability $\binom{d}{2}/\binom{2m}{2}$ is about 1/2 which occurs for the stub proportion close to $1/\sqrt{2} \approx 0.7$.

The asymptotic entropies (Frank and Nowicki 1989) of the loop multiplicity distribution under RSM and under IEA are obtained by normal approximations given by

$$h(\mathbf{P}) \approx \frac{1}{2} \log [2\pi e(\sigma_{ii}^2 + \Delta_{ii})] ,$$

and

$$h(\mathbf{B}) \approx \frac{1}{2} \log [2\pi e\sigma_{ii}^2] .$$

For the case where $m = 10$ and $d = 10$, these approximations are equal to $h(\mathbf{P}) \approx 1.747$ and $h(\mathbf{B}) \approx 2.474$. Figure 5 shows how these entropy approximations (dotted lines) vary for the same cases as in Figure 4, i.e. for different stub proportions $d/2m$ for $m = 20, 40, 60$. We see that the approximations are close to their true values for all cases shown.

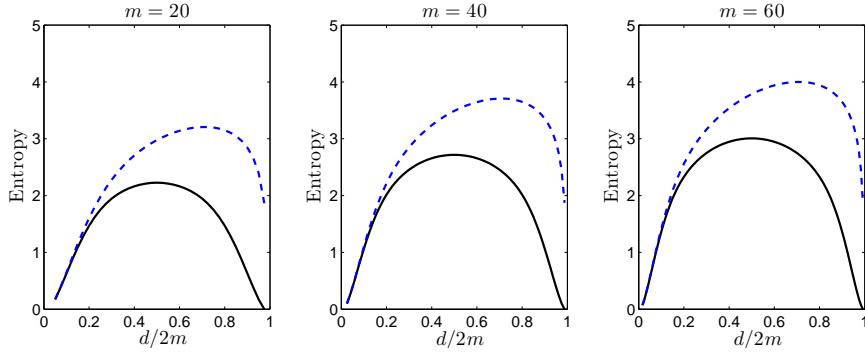


Figure 4: Entropy of the distribution of loop multiplicity under random stub matching (solid lines) and independent edge assignments (dashed lines) at a vertex of degree d in a graph with m edges. Entropy is plotted against stub proportion $d/2m$ for $m = 20, 40, 60$.

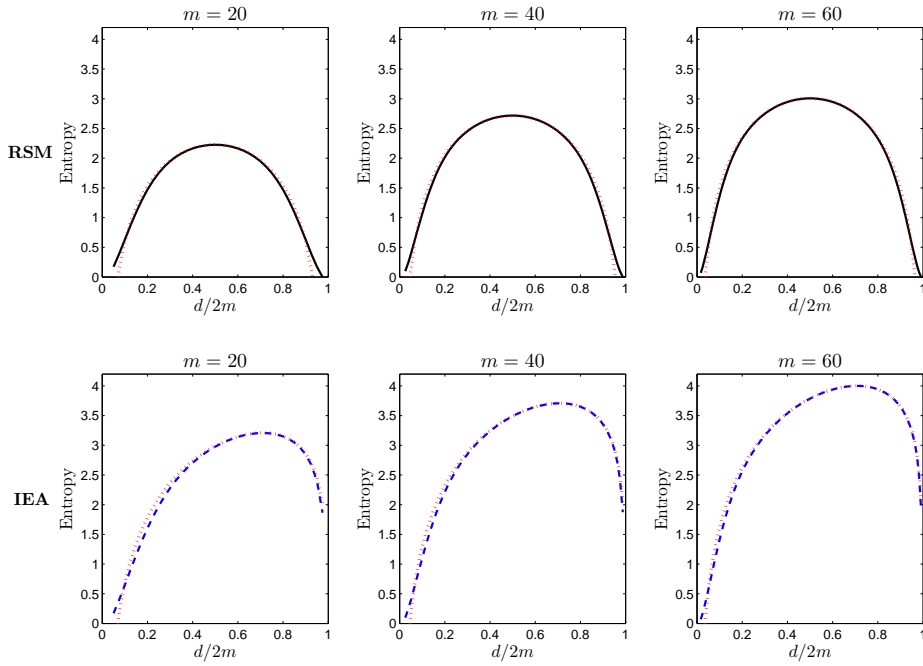


Figure 5: Entropy of the distribution of loop multiplicity under random stub matching (solid lines) and independent edge assignments (dashed lines) at a vertex of degree d in a graph with m edges. The entropy approximations are illustrated with dotted lines for all cases. Entropy is plotted against stub proportion $d/2m$ for $m = 20, 40, 60$.

We now turn to edge multiplicities $m_{ij} = m_{ij}(\boldsymbol{\eta})$ for $i < j$ under RSM, and start with the joint probability distribution of the multiplicities $(m_{ii}, m_{jj}, m_{ij}) = (m_{ii}(\boldsymbol{\eta}), m_{jj}(\boldsymbol{\eta}), m_{ij}(\boldsymbol{\eta}))$ which is denoted $P_{uvw} = P((m_{ii}, m_{jj}, m_{ij}) = (u, v, w))$. Applying a similar argument as before for i -loops, j -loops, (i, j) -edges, remaining i -stubs and j -stubs, we obtain after simplification the following formula for the trivariate probabilities under RSM:

$$P_{uvw} = \frac{\binom{m}{u, v, w, d_i - 2u - w, d_j - 2v - w} 2^{d_i + d_j - 2u - 2v - w}}{\binom{2m}{d_i, d_j}}.$$

To specify possible outcomes of (u, v, w) let a and b denote the smallest and largest number of stubs at vertices i and j , and c denote the number of stubs at other vertices, i.e. $c = 2m - a - b$. Table 4 gives the number of stubs of each category occurring at the edges within and between categories. There are u loops with $2u$ stubs at the vertex with a stubs, v loops with $2v$ stubs at the vertex with b stubs, and w edges with w stubs of each kind between these two vertices. There are $a - 2u - w$ and $b - 2v - w$ remaining stubs at these two vertices, and they combine to edges with the same number of other stubs. Since there is a total of c other stubs, there remain $c - (a - 2u - w) - (b - 2v - w) = 2(m - a - b + u + v + w)$ other stubs giving room for $m - a - b + u + v + w$ other loops or edges. Note that $2u + w \leq a$, $2v + w \leq b$ and $(2u + w) + (2v + w) \geq a + b - c$ are required to achieve non-negative frequencies.

Table 4: Number of stubs of each vertex category at edges or loops within and between categories.

	Vertex i	Vertex j	Other vertices	Total
Vertex i	$2u$	w	$a - 2u - w$	a
Vertex j	w	$2v$	$b - 2v - w$	b
Other vertices	$a - 2u - w$	$b - 2v - w$	$2(m - a - b + u + v + w)$	c
Total	a	b	c	$2m$

The set of possible outcomes (u, v, w) is illustrated in Figure 6 and correspond to marked points in the shaded regions with possible stub frequencies $(2u + w, 2v + w)$ of the same parity. Four different cases are numerically illustrated in Figure 6 where $(a, b) = (3, 7)$ and the value of c is varied and given by the vertical distance from $a + b$ to the upper point of the digonal. First, consider $b \leq a + b - c \leq a + b$ which corresponds to $0 \leq c \leq a$ and choose $c = 2$. There are four possible points in the triangular region, namely $(1, 7)$, $(2, 6)$,

$(3, 5), (3, 7)$ corresponding to (u, v, w) equal to $(0, 3, 1), (1, 3, 0), (0, 2, 2), (1, 2, 1), (0, 1, 3), (1, 3, 1), (0, 2, 3)$. These (u, v, w) are obtained by choosing $w = 0, 2, \dots$ or $w = 1, 3, \dots$ so that $2u$ and $2v$ get even. It can be shown that the possible point $(2u + w, 2v + w)$ in the shaded region corresponds to

$$1 + \left\lfloor \frac{\min(2u + w, 2v + w)}{2} \right\rfloor$$

possible outcomes (u, v, w) . Second, consider the case where $a \leq a + b - c \leq b$ which corresponds to $a \leq c \leq b$ and choose $c = 6$. Third, consider $0 \leq a + b - c \leq a$ which corresponds to $a \leq b \leq c \leq a + b$ and choose $c = 8$, and finally fourth, consider $a + b - c \leq 0$ which corresponds to $a + b \leq c$ and choose $c = 12$ to illustrate the case with maximal number of outcomes.

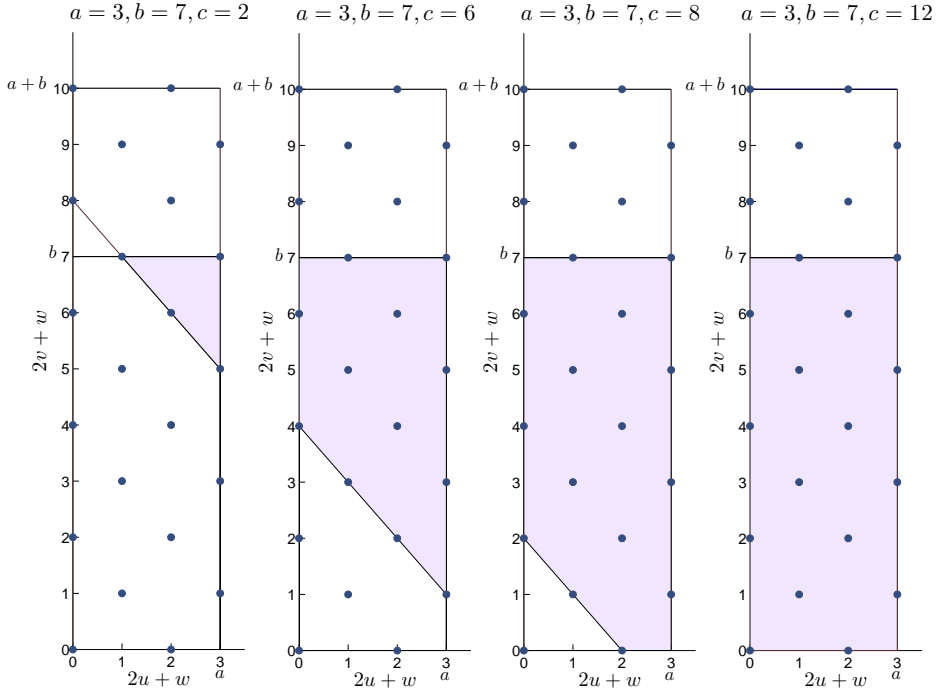


Figure 6: The possible outcomes of (u, v, w) in Table 4 correspond to the shaded region with stub frequencies $(2u + w, 2v + w)$ of the same parity. The four cases illustrate how the region varies when the number of other stubs c is smaller than a , between a and b , between b and $a + b$, and larger than $a + b$.

Letting $u \leq \lfloor (a-w)/2 \rfloor = \alpha_w$, $v \leq \lfloor (b-w)/2 \rfloor = \beta_w$, and $u+v \geq a+b-m-w = \gamma_w$, the total number of possible outcomes of (u, v, w) denoted K is given by

$$K = \sum_{w=0}^a K_w ,$$

where

$$K_w = \begin{cases} \binom{\alpha_w + \beta_w - \gamma_w + 2}{2} & \text{if } \gamma_w \geq \beta_w \\ (\alpha_w + 1)(\beta_w - \gamma_w) + \binom{\alpha_w + 2}{2} & \text{if } \alpha_w \leq \gamma_w \leq \beta_w \\ (\alpha_w + 1)(\beta_w + 1) - \binom{\gamma_w + 1}{2} & \text{if } \gamma_w \leq \alpha_w , \end{cases}$$

providing an upper bound to the entropy. Table 5 gives a numerical example of this result for the second case in Figure 6, i.e. when $a = 3$, $b = 7$ and $c = 6$. The twelve points in Figure 6 for this case can be individually checked for possible (u, v, w) . There are six points with two outcomes and six points with one outcome of (u, v, w) , thus a total of 18 outcomes. Using the formula for K_w we also find that the total number of possible outcomes here is given by

$$K = \sum_{w=0}^3 K_w = 18 ,$$

implying that the entropy of this distribution is

$$h(\mathbf{P}) = \sum_{uvw: P_{uvw} > 0} -P_{uvw} \log P_{uvw} \leq \log(18) = 4.170 ,$$

where $\mathbf{P} = (P_{uvw} : \text{all possible } (u, v, w))$. The exact entropy in this case turns out to be 3.525. Further comparisons are presented in Table 6.

Table 5: The total number of possible outcomes of (u, v, w) in Table 4 when $a = 3$, $b = 7$ and $c = 6$, where $u \leq \alpha_w$, $v \leq \beta_w$, and $u + v \geq \gamma_w$.

w	α_w	β_w	γ_w	K_w
0	1	3	2	5
1	1	3	1	7
2	0	2	0	3
3	0	2	-1	3
				$K = 18$

If the edges are assumed to be independently assigned to sites, the IEA distribution for (m_{ii}, m_{jj}, m_{ij}) is multinomial distribution with parameters m and $[Q_{ii} \ Q_{jj} \ Q_{ij} \ (1 - Q_{ii} - Q_{jj} - Q_{ij})]$ where Q_{ij} for $i \leq j$ is defined as earlier. This distribution is shortly denoted \mathbf{B} with probabilities B_{uvw} for $\binom{m+3}{3}$ different outcomes $(u, v, w, m - u - v - w)$ that are ordered partitions of m into four non-negative integers. Thus, the entropy of this distribution is

$$h(\mathbf{B}) = \sum_{uvw: B_{uvw} > 0} -B_{uvw} \log B_{uvw} \leq \log \binom{m+3}{3}$$

and using the normal approximation to the multinomial distribution we obtain the approximate entropy

$$h(\mathbf{B}) \approx \frac{1}{2} \log [(2\pi e)^3 \det(\boldsymbol{\Sigma}_{\text{IEA}})] ,$$

where $\boldsymbol{\Sigma}_{\text{IEA}}$ is the covariance matrix of (m_{ii}, m_{jj}, m_{ij}) under IEA given by

$$\boldsymbol{\Sigma}_{\text{IEA}} = m \begin{pmatrix} Q_{ii}(1 - Q_{ii}) & -Q_{ii}Q_{jj} & -Q_{ii}Q_{ij} \\ -Q_{jj}Q_{ii} & Q_{jj}(1 - Q_{jj}) & -Q_{jj}Q_{ij} \\ -Q_{ij}Q_{ii} & -Q_{ij}Q_{jj} & Q_{ij}(1 - Q_{ij}) \end{pmatrix} .$$

The determinant of $\boldsymbol{\Sigma}_{\text{IEA}}$ is given by

$$\begin{aligned} \det(\boldsymbol{\Sigma}_{\text{IEA}}) &= m^3 Q_{ii} Q_{jj} Q_{ij} [(1 - Q_{ii})(1 - Q_{jj})(1 - Q_{ij}) - 2Q_{ii}Q_{jj}Q_{ij} \\ &\quad - (1 - Q_{ii})Q_{jj}Q_{ij} - (1 - Q_{jj})Q_{ii}Q_{ij} - (1 - Q_{ij})Q_{ii}Q_{jj}] \\ &= m^3 Q_{ii} Q_{jj} Q_{ij} (1 - Q_{ii} - Q_{jj} - Q_{ij}) , \end{aligned}$$

so that

$$h(\mathbf{B}) \approx \frac{1}{2} \log [(2\pi e m)^3 Q_{ii} Q_{jj} Q_{ij} (1 - Q_{ii} - Q_{jj} - Q_{ij})] .$$

The approximate entropy of the distribution of (m_{ii}, m_{jj}, m_{ij}) under RSM is

$$h(\mathbf{P}) \approx \frac{1}{2} \log [(2\pi e)^3 \det(\boldsymbol{\Sigma}_{\text{RSM}})] ,$$

where $\det(\boldsymbol{\Sigma}_{\text{RSM}})$ is the determinant of the covariance matrix. The covariance matrix of (m_{ii}, m_{jj}, m_{ij}) under RSM is given by

$$\boldsymbol{\Sigma}_{\text{RSM}} = \boldsymbol{\Sigma}_{\text{IEA}} + \boldsymbol{\Delta} ,$$

where

$$\boldsymbol{\Delta} = m(m-1) \begin{pmatrix} Q_{iii} - Q_{ii}^2 & Q_{iij} - Q_{ii}Q_{jj} & Q_{iij} - Q_{ii}Q_{ij} \\ Q_{iij} - Q_{ii}Q_{jj} & Q_{jjj} - Q_{jj}^2 & Q_{ijj} - Q_{ij}Q_{jj} \\ Q_{iij} - Q_{ii}Q_{ij} & Q_{ijj} - Q_{ij}Q_{jj} & Q_{ijj} - Q_{ij}^2 \end{pmatrix} ,$$

with

$$Q_{iiii} = \frac{\binom{d_i}{2} \binom{d_i-2}{2}}{\binom{2m}{2} \binom{2m-2}{2}}, \quad Q_{iijj} = \frac{\binom{d_i}{2} \binom{d_j}{2}}{\binom{2m}{2} \binom{2m-2}{2}},$$

$$Q_{iiij} = \frac{\binom{d_i}{2} (d_i-2) d_j}{\binom{2m}{2} \binom{2m-2}{2}}, \quad Q_{ijij} = \frac{d_i (d_i-1) d_j (d_j-1)}{\binom{2m}{2} \binom{2m-2}{2}},$$

and note that $Q_{ijkl} = Q_{klij}$ for all $i \leq j$ and $k \leq \ell$.

Table 6 illustrates the different entropies presented here for some cases with $a = 3$, $b = 7$ and where the total edge frequency m is varied, including the four cases given in Figure 6. Also presented in this table is the information divergence between the RSM and IEA distribution of (m_{ii}, m_{jj}, m_{ij}) :

$$D(\mathbf{P}, \mathbf{B}) = \sum_{uvw: P_{uvw} > 0} P_{uvw} \log \frac{P_{uvw}}{B_{uvw}}.$$

We see in Table 6 that the approximate entropies are close to the entropies of both the IEA and RSM distributions indicating that both distributions are fairly well approximated by the normal distribution. We also note that as m increases, the RSM entropy moves towards that of the IEA. This can also be seen by the divergence values which are decreasing towards zero for increasing m . The common limiting distributions of RSM and IEA is a one-point distribution at $(u, v, w) = (0, 0, 0)$ so the exact entropies tend to zero. The upper bounds for the RSM distributions tend to $\log(22)$ since there are 22 points (u, v, w) corresponding to the last case shown in Figure 6. This limit is achieved already for $m = 10$.

The distribution of a single non-loop multiplicity $m_{ij} = m_{ij}(\boldsymbol{\eta})$ for $i < j$ under RSM is given as a marginal in the trivariate distribution of (m_{ii}, m_{jj}, m_{ij}) . It is obtained by summing over the numbers of loops at vertices i and j . Thus,

$$P(m_{ij} = w) = P_{..w} = \sum_{u=0}^{\lfloor \frac{a-w}{2} \rfloor} \sum_{v=0}^{\lfloor \frac{b-w}{2} \rfloor} P_{uvw}, \quad \text{for } w = 0, 1, \dots, a,$$

where $a = \min(d_i, d_j) \geq 1$ and $b = \max(d_i, d_j)$ with $a + b \leq 2m$. Note that not all $P_{uvw} > 0$. For the special case when $n = 2$, $a + b = 2m$ and $u + v + w = m$, we get $a = 2u - w$ and $b = 2v - w$ and the marginal distribution of m_{12} simplifies to

$$P(m_{12} = w) = P_w = \frac{\binom{m}{u, v, w} 2^w}{\binom{2m}{a}} = \frac{m! 2^w a! (2m-a)!}{(2m)! \left(\frac{a-w}{2}\right)! \left(\frac{b-w}{2}\right)! w!},$$

where $w = 0, 2, \dots, a$ if a and b are even, and $w = 1, 3, \dots, a$ if a and b are odd. For this case we cannot expect the binomial distribution under IEA to be an adequate approximation. Obviously the IEA distribution $\mathbf{B} = (B_w : w = 0, 1, \dots, m)$ with parameters m and

Table 6: Entropy of the joint edge multiplicity distribution under random stub matching (RSM), independent edge assignments (IEA) and the entropy approximations for these distributions where number of stubs are $a = 3$, $b = 7$ and the total edge frequency is $m = 6, 8, 9, 11, 20, 30, 40, 50, 60$, thus including the four cases shown in Figure 6. Also given is the divergence between these two distributions.

m	Entropy RSM			Entropy IEA			Divergence
	Upper bound	Exact	Approximate	Upper bound	Exact	Approximate	
6	2.81	2.36	2.31	6.39	5.07	5.31	1.88
8	4.17	3.53	3.61	7.37	4.87	5.13	0.76
9	4.39	3.64	3.70	7.78	4.68	4.94	0.55
11	4.46	3.61	3.65	8.51	4.30	4.58	0.34
20	4.46	2.96	2.90	10.79	3.17	3.36	0.10
30	4.46	2.40	2.19	12.41	2.50	2.48	0.04
40	4.46	2.03	1.65	13.59	2.08	1.86	0.02
50	4.46	1.77	1.21	14.52	1.80	1.38	0.02
60	4.46	1.57	0.84	15.28	1.59	0.10	0.01

$ab/\binom{2m}{2} = a(2m - a)/m(2m - 1)$ gives positive probabilities to all outcomes whereas the RSM distribution of edge multiplicity, $\mathbf{P} = (P_w : w = 0, 1, \dots, m)$, has zero probabilities for all even or all odd outcomes. According to the results in Section 4, it is only for this special case, $n = 2$, that the RSM distribution can have a variance $\sigma_{ij}^2 + \Delta_{ij}$ that is larger than the variance σ_{ij}^2 of the IEA distribution. 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 σ_{ij}^2 , so even if the number of cases increases with increasing m , the proportion of cases decreases towards zero. This is illustrated in Figure 7, where we also notice that the proportion is not monotonically decreasing.

Table 7 gives the RSM distribution of edge multiplicity and the corresponding IEA distribution for the case $m = 10$ and $(a, 2m - a) = (10, 10)$. Also presented in Table 7 is the information divergence between these distributions. The entropies for this example are equal to $h(\mathbf{P}) = 1.746$ and $h(\mathbf{B}) = 2.704$. The asymptotic entropies for the edge multiplicity

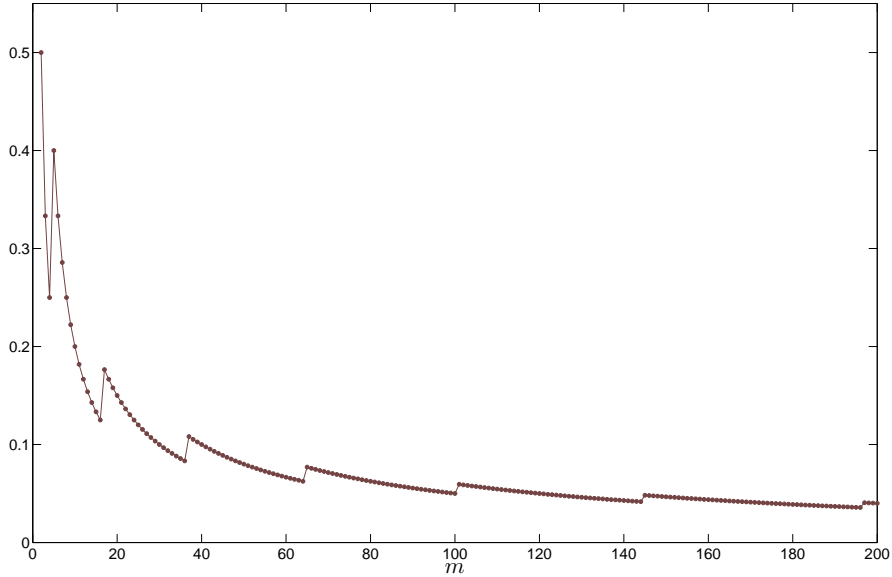


Figure 7: Proportion of the degree pairs $(a, 2m - a)$ for $a = 1, \dots, m$ with edge multiplicity variance larger for random stub matching than for independent edge assignments. The proportions are plotted against edge frequency m .

distributions under RSM and IEA give the following approximations:

$$h(\mathbf{P}) \approx \frac{1}{2} \log [2\pi e(\sigma_{ij}^2 + \Delta_{ij})]$$

and

$$h(\mathbf{B}) \approx \frac{1}{2} \log [2\pi e\sigma_{ij}^2] .$$

For the case where $m = 10$ and $(a, 2m - a) = (10, 10)$, these approximations are equal to $h(\mathbf{P}) \approx 2.747$ and $h(\mathbf{B}) \approx 2.706$.

Figure 8 shows divergence for m_{ij} at degree pairs $(a, 2m - a)$ for different a when $m = 10$, and how it varies for different proportions a/m for some selected values of m . Figure 9 highlights how $h(\mathbf{P})$ and $h(\mathbf{B})$ vary for different a when m and Figure 10 compares the entropy approximations to their true values. The deviation between entropy values in Figure 10 indicates that edge multiplicity under RSM is poorly approximated by a normal distribution. However, the approximate entropies are close to the true values under IEA.

Table 7: The probability distribution of edge multiplicity at a pair of vertices with degree pair $(a, 2m - a) = (10, 10)$ when $m = 10$ edges are formed by random stub matching (RSM). It is compared by information divergence to the binomial distribution obtained by independent edge assignments (IEA).

Number of edges	Probability under RSM	Probability under IEA	Weighted log-likelihood ratio
0	0.001364	0.000569	0.001721
1	0	0.006319	0
2	0.068198	0.031595	0.075702
3	0	0.093614	0
4	0.363723	0.182028	0.363242
5	0	0.242704	0
6	0.436468	0.224726	0.418008
7	0	0.142683	0
8	0.1247050	0.059451	0.133277
9	0	0.014679	0
10	0.0055424	0.001631	0.009781
			Divergence = 1.001733

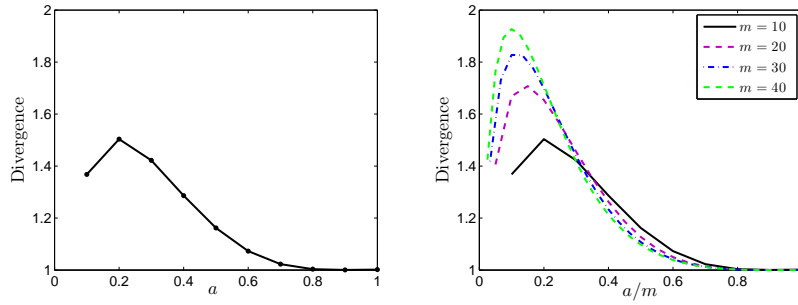


Figure 8: Information divergence between random stub matching and independent edge assignments for the distributions of edge multiplicity at a pair of vertices with degree pair $(a, 2m - a)$ in a graph with m edges. Information divergence is plotted against different degrees a for $m = 10$ and against different proportions a/m for $m = 10, 20, 30, 40$.

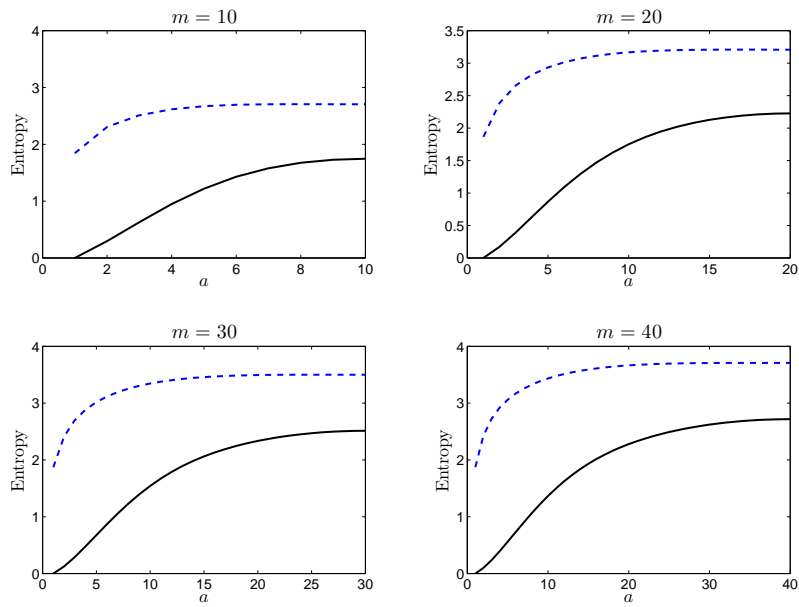


Figure 9: Entropy of the distribution of edge multiplicity under random stub matching (solid lines) and independent edge assignments (dashed lines) at a pair of vertices with degree pair $(a, 2m - a)$ in a graph with m edges. Entropy is plotted against different degrees a for $m = 10, 20, 30, 40$.

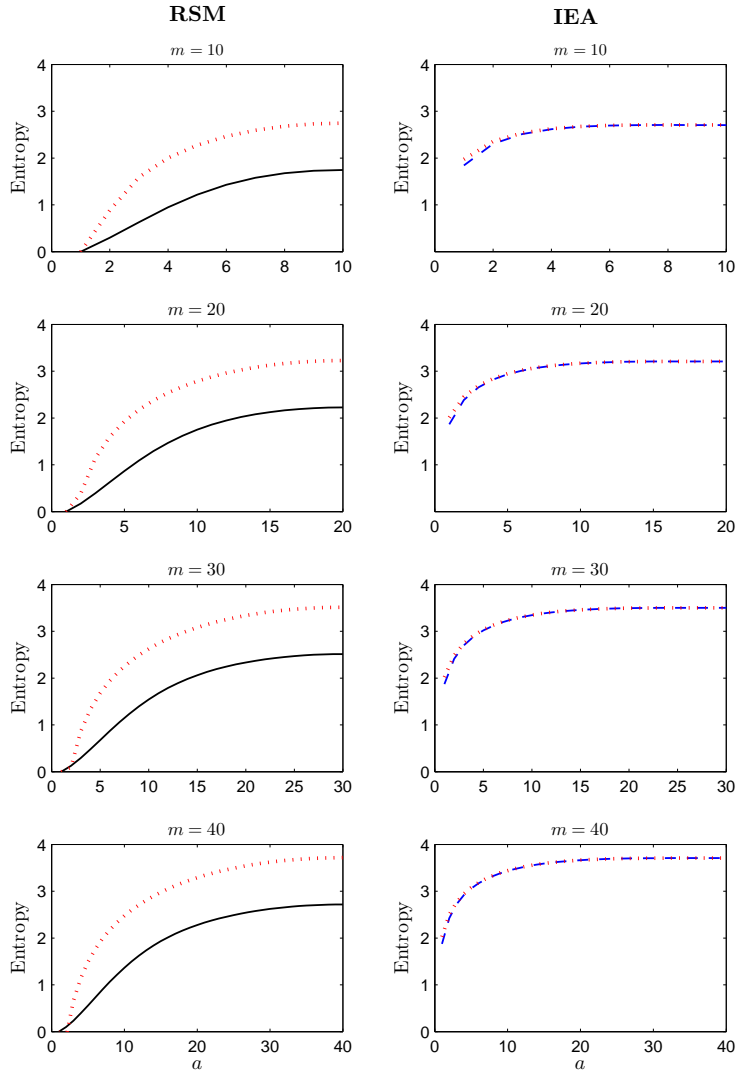


Figure 10: Entropy of the distribution of edge multiplicity under random stub matching (solid lines) and independent edge assignments (dashed lines) at a pair of vertices with degree pair $(a, 2m - a)$ in a graph with m edges. The entropy approximations are illustrated with dotted lines for all cases. Entropy is plotted against different degrees a for $m = 10, 20, 30, 40$.

This section is concluded 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 $m = 15$ is illustrated in Figure 11 where divergence $D(\mathbf{P}, \mathbf{B})$ is plotted against degree pairs (a, b) using a color coding of standardized divergence values applied to the unit squares located around points (a, b) . The divergences for all possible degree pairs (a, b) are calculated and their maxima are determined. Standardized divergence values are obtained by dividing with the maxima. Every 10th percentile of this standardized distribution is then calculated and assigned a color where darker colors represent higher divergences, i.e. darker colors 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 11 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 11 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 11 represents cases where $a \leq c \leq b$. Here, the majority of the region has a brighter color implying less deviation between the RSM distribution and the IEA distribution. The same applies for the lower region in Figure 11 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.

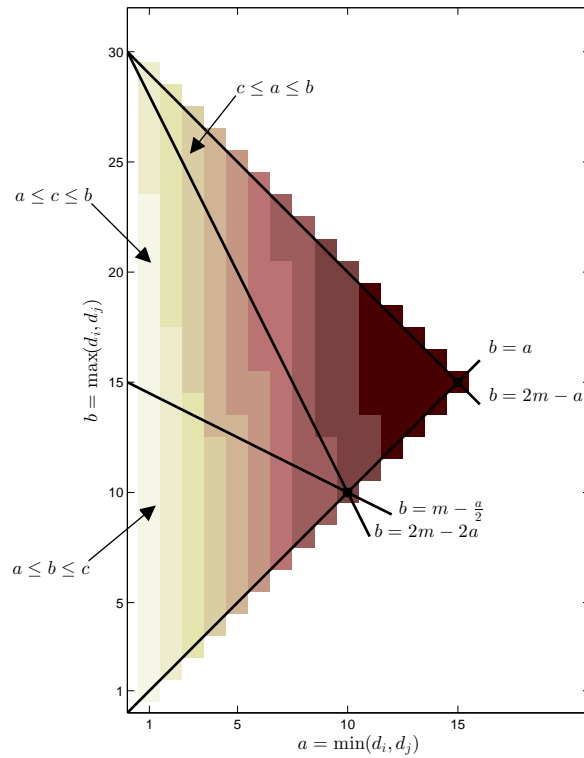


Figure 11: Divergence between the multiplicity distribution under random stub matching and under independent edge assignments for edges between two vertices with ordered degree pair (a, b) when total edge frequency is $m = 15$ and total number of stubs is $2m = a + b + c$. A darker color at the unit squares located around points (a, b) represent a larger divergence than a brighter color.

6 Distributions of Multigraphs

So far we have been considering the local structure of multigraphs by examining the distribution of edge multiplicities $m_{ij}(\boldsymbol{\eta})$ for $i \leq j$. We now turn to the global structure by studying the distribution of multigraphs, i.e. the distribution of $\mathbf{m}(\boldsymbol{\eta})$. Entropy measures are used to compare the distributions of multigraphs under RSM and under IEA.

In Section 3, the probability of generating undirected multigraphs under RSM was shown to be given by

$$P(\mathbf{m}(\boldsymbol{\eta}) = \mathbf{m}) = P_{\mathbf{m}}(\mathbf{d}) = \frac{2^{m_2} \binom{m}{\mathbf{m}}}{\binom{2m}{\mathbf{d}}} .$$

This probability depends on a single statistic, the complexity measure $t(\boldsymbol{\eta}) = m_1(\boldsymbol{\eta}) + \log \mathbf{m}(\boldsymbol{\eta})!$, and can be written as

$$P_{\mathbf{m}}(\mathbf{d}) = C 2^{-t} ,$$

where $t = m_1 + \log \mathbf{m}!$ is the outcome of $t(\boldsymbol{\eta})$ and $C = 2^m m! \mathbf{d}! / (2m)!$ is a constant. Letting $\mathbf{P} = (P_{\mathbf{m}}(\mathbf{d}) : \text{all different } \mathbf{m})$ denote the probability distribution of multigraphs under RSM, its entropy $H(\mathbf{m}(\boldsymbol{\eta})) = h(\mathbf{P})$ is given by

$$\begin{aligned} h(\mathbf{P}) &= \sum_{\mathbf{m}: P_{\mathbf{m}} > 0} -P_{\mathbf{m}} \log P_{\mathbf{m}} \\ &= E(t(\boldsymbol{\eta})) - \log C , \end{aligned}$$

where the expected value of $t(\boldsymbol{\eta})$ is

$$E(t(\boldsymbol{\eta})) = E(m_1) + E(\log \mathbf{m}!) .$$

The first term of the above sum is the expected number of loops $m_1 = m_1(\boldsymbol{\eta})$ under RSM which by using the results in Section 4 is obtained as

$$E(m_1) = \sum_{i=1}^n E(m_{ii}) = m \sum_{i=1}^n Q_{ii} = \frac{1}{2m-1} \sum_{i=1}^n \binom{d_i}{2} .$$

The second term of the expression for the expected value of $t(\boldsymbol{\eta})$ can be expanded to

$$E(\log \mathbf{m}!) = \sum_{k=2}^m \log k! \sum_{i \leq j} \sum P(m_{ij} = k) ,$$

where the probabilities need to be considered for vertex pairs with different degree pairs. Letting $P(m_{ii} = k) = P_k(m, a)$ for $d_i = a$, $P(m_{ij} = k) = P_k(m, a, b)$ for $a = \min(d_i, d_j)$ and $b = \max(d_i, d_j)$, $\sum_{i=1}^n I(d_i = a) = n_a$, $\sum_{i=1}^n I(d_i = b) = n_b$ and

$$\sum_{i < j} \sum I(\min(d_i, d_j) = a, \max(d_i, d_j) = b) = \begin{cases} n_a n_b & \text{for } a < b \\ \binom{n_a}{2} & \text{for } a = b , \end{cases}$$

we have that

$$\sum_{i \leq j} \sum P(m_{ij} = k) = \sum_a n_a P_k(m, a) + \sum_a \binom{n_a}{2} P_k(m, a, a) + \sum_{a < b} \sum n_a n_b P_k(m, a, b) .$$

These expansions yield the formula for the exact entropy. In particular, for regular graphs with the same degree $d = 2m/n$ at each vertex, the expected number of loops under RSM is given by

$$E(m_1) = \frac{n}{(2m-1)} \binom{d}{2} = \frac{nd(d-1)}{2(nd-1)} ,$$

and the exact entropy is simplified to

$$h(\mathbf{P}) = \frac{nd(d-1)}{2(nd-1)} + \sum_{k=2}^m \log k! \left[n P_k(m, d) + \binom{n}{2} P_k(m, d, d) \right] - \log C ,$$

where

$$C = \frac{(d!)^n}{(nd-1)!!} .$$

The approximate entropy of the distribution of multigraphs under RSM is given by

$$h(\mathbf{P}) \approx \frac{1}{2} \log \left[(2\pi e)^{r-n} \det(\boldsymbol{\Sigma}_{\text{RSM}}) \right] ,$$

where $\boldsymbol{\Sigma}_{\text{RSM}}$ is the covariance matrix of $r - n$ non-redundant components of the multiplicity sequence \mathbf{m} . In order to find $\boldsymbol{\Sigma}_{\text{RSM}}$, consider Q_{ijkl} for all $i \leq j$ and $k \leq \ell$ where $Q_{ijkl} = Q_{klij}$. More specifically, we need the formulae for two loops that are at the same vertex or at different vertices, one loop and one non-loop with one or no common vertex, and two non-loops with two, one or no vertices in common. With a slight abuse of notation, the r by r covariance matrix under RSM can be written as

$$\boldsymbol{\Sigma}_{\text{RSM}} = \boldsymbol{\Sigma}_{\text{IEA}} + \boldsymbol{\Delta} ,$$

where $\boldsymbol{\Sigma}_{\text{IEA}}$ is the r by r covariance matrix under independent edge assignments, i.e. the covariance matrix of a multinomial distribution with parameters m and \mathbf{Q} . The elements of $\boldsymbol{\Sigma}_{\text{IEA}}$ are $m Q_{ij} (\delta_{ijkl} - Q_{kl})$ for r different (i, j) and (k, ℓ) in R , where δ_{ijkl} is equal to 1 if $(i, j) = (k, \ell)$ and 0 otherwise. The matrix $\boldsymbol{\Delta}$ consists of elements $\Delta_{ijkl} = m(m-1)(Q_{ijkl} - Q_{ij}Q_{kl})$ for r different (i, j) and (k, ℓ) in site space R . Renaming the ordered edge sequence indexes $(1, 1) < (1, 2) < \dots < (1, n) < (2, 2) < (2, 3) < \dots < (n, n)$ to $1, 2, \dots, r = \binom{n+1}{2}$, $\boldsymbol{\Sigma}_{\text{RSM}}$ can be written as

$$\boldsymbol{\Sigma}_{\text{RSM}} = m \begin{pmatrix} Q_1(1-Q_1) & -Q_1Q_2 & \cdots & -Q_1Q_r \\ -Q_1Q_2 & Q_2(1-Q_2) & \cdots & -Q_2Q_r \\ \vdots & \vdots & \ddots & \vdots \\ -Q_1Q_r & -Q_2Q_r & \cdots & Q_r(1-Q_r) \end{pmatrix} + m(m-1) \begin{pmatrix} \Delta_{11} & \Delta_{12} & \cdots & \Delta_{1r} \\ \Delta_{12} & \Delta_{22} & \cdots & \Delta_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{1r} & \Delta_{2r} & \cdots & \Delta_{rr} \end{pmatrix} .$$

In order to avoid singularity of Σ_{RSM} , remove n of the r components in \mathbf{m} that are linear combinations of the others, implying that the degrees of freedom here is equal to $r - n = \binom{n}{2}$. A similar argument was used in Section 5 for the trivariate distribution of the edge multiplicities (m_{ii}, m_{jj}, m_{ij}) obtained when three of the six pairs of vertex categories were redundant.

We now turn to the distribution of multigraphs under IEA which corresponds to that edges in ξ are independently assigned to sites according to the probability distribution $(P_{ij} : (i, j) \in V^2)$, and edges in η are independently assigned to sites according to the probability distribution $\mathbf{Q} = (Q_{ij} : (i, j) \in R)$, where P_{ij} and Q_{ij} are defined as earlier. Thus,

$$P(\boldsymbol{\eta} = \mathbf{y}) = \prod_{k=1}^m Q(y_{2k-1}, y_{2k}) = \prod_{i \leq j} Q_{ij}^{m_{ij}(\mathbf{y})} ,$$

where $Q(i, j) = Q_{ij}$, and $\mathbf{m}(\boldsymbol{\eta})$ is multinomially distributed with parameters m and \mathbf{Q} so that

$$P(\mathbf{m}(\boldsymbol{\eta}) = \mathbf{m}) = \binom{m}{\mathbf{m}} \mathbf{Q}^{\mathbf{m}} = \frac{m!}{\prod_{i \leq j} m_{ij}!} \prod_{i \leq j} Q_{ij}^{m_{ij}} ,$$

which is shortly denoted $B_{\mathbf{m}}(\mathbf{d})$. Note that this implies that $m_{ij}(\boldsymbol{\eta})$ is binomially distributed with parameters m and Q_{ij} , as considered in previous sections. Letting $\mathbf{B} = (B_{\mathbf{m}}(\mathbf{d}) : \text{all different } \mathbf{m})$ denote the probability distribution of multigraphs under IEA, the entropy of this distribution is equal to

$$h(\mathbf{B}) = \sum_{\mathbf{m}: B_{\mathbf{m}} > 0} -B_{\mathbf{m}} \log B_{\mathbf{m}} .$$

The number of multigraphs \mathbf{m} is not restricted by \mathbf{d} in the same way as for the RSM distribution. Here it is given by the total number of ordered partitions of m into $r = \binom{n+1}{2}$ available sites of vertex pairs for the edges:

$$\#\mathbf{m} = \binom{m + r - 1}{m} .$$

This gives an upper bound to the entropy

$$h(\mathbf{B}) \leq \log \binom{m + r - 1}{m} .$$

The approximate entropy under IEA is given by

$$h(\mathbf{B}) \approx \frac{1}{2} \log [(2\pi e)^{r-1} \det(\Sigma_{\text{IEA}})] ,$$

where Σ_{IEA} now denotes the $(r - 1)$ by $(r - 1)$ covariance matrix of \mathbf{m} when one of the r components is omitted (to avoid singularity). The determinant of this matrix can be proved to be equal to

$$m^{r-1} \prod_{i \leq j} Q_{ij} ,$$

where the product is over all r pairs $(i, j) \in R$. Thus, the approximate entropy under IEA is obtained by

$$h(\mathbf{B}) \approx \frac{1}{2} \log \left[(2\pi em)^{r-1} \prod_{i \leq j} Q_{ij} \right] .$$

In Table 8 we consider the entropies of the distributions of multigraphs under RSM and IEA and the entropy approximations for these distributions with $n = 6$ vertices, $m = 9$ edges and different degree sequences with minimum degree 2. Also shown in Table 8 is the divergence between these two distributions. The total number of graphs under IEA for this example is 10,015,005 and we do not calculate the exact entropies which here are close to their approximations. We see that the approximate entropies under RSM and under IEA are both close to their upper bounds. The exact entropies under RSM are close to the approximate entropies implying that the distributions of multigraphs are fairly well approximated by normal distributions. However, the divergence values indicate a large deviation between the distributions under RSM and IEA. These findings indicate rather flat distributions over very different ranges for RSM and IEA. We note that there are cases in Table 8 when the approximate entropies are larger than the upper bounds to the exact entropies. This occurs for the last three rows of Table 8. It would be a tedious task to investigate this in general by using the formula given in Section 5 for the total number of multigraphs under RSM. We therefore restrict ourselves to a general investigation of IEA by considering the following inequality which holds when the IEA entropy approximation is at most equal to the upper bound to the exact entropy:

$$(2\pi em)^{\frac{r-1}{2}} \left[\prod_{i \leq j} Q_{ij} \right]^{\frac{1}{2}} \leq \binom{m+r-1}{m} .$$

The right hand side can be written as

$$\binom{m+r-1}{m} = \prod_{k=1}^{r-1} \left(1 + \frac{m}{k} \right) ,$$

and the inequality can be expressed as giving an upper bound to the geometric mean of the edge probabilities according to

$$\tilde{Q} \leq \left(\frac{G}{2\pi e} \right)^{\frac{r-1}{r}} ,$$

Table 8: Entropies of the distributions of multigraphs under random stub matching (RSM), independent edge assignments (IEA) and the entropy approximations for these distributions with $n = 6$ vertices, $m = 9$ edges and different degree sequences with minimum degree 2. Also given is the divergence between these two distributions.

Degree sequence	Number of graphs	Entropy RSM			Entropy IEA		Divergence
		Upper bound	Exact	Approx	Upper bound	Approx	
$\mathbf{d} = (8, 2, 2, 2, 2, 2)$	773	9.59	8.63	7.61	23.26	18.84	8.72
$\mathbf{d} = (7, 3, 2, 2, 2, 2)$	1210	10.24	9.47	8.98	23.26	20.41	8.90
$\mathbf{d} = (6, 4, 2, 2, 2, 2)$	1651	10.69	9.97	9.62	23.26	21.15	8.99
$\mathbf{d} = (5, 5, 2, 2, 2, 2)$	1804	10.82	10.14	9.82	23.26	21.37	9.02
$\mathbf{d} = (6, 3, 3, 2, 2, 2)$	1914	10.90	10.23	10.25	23.26	21.86	9.07
$\mathbf{d} = (5, 4, 3, 2, 2, 2)$	2424	11.24	10.60	10.77	23.26	22.45	9.14
$\mathbf{d} = (4, 4, 4, 2, 2, 2)$	2814	11.46	10.81	11.08	23.26	22.82	9.18
$\mathbf{d} = (5, 3, 3, 3, 2, 2)$	2857	11.48	10.87	11.40	23.26	23.17	9.21
$\mathbf{d} = (4, 4, 3, 3, 2, 2)$	3316	11.70	11.08	11.72	23.26	23.53	9.26
$\mathbf{d} = (4, 3, 3, 3, 3, 2)$	3943	11.95	11.36	12.35	23.26	23.27	9.33
$\mathbf{d} = (3, 3, 3, 3, 3, 3)$	4720	12.21	11.64	12.98	23.26	24.97	9.41

where

$$\tilde{Q} = \left[\prod_{i \leq j} Q_{ij} \right]^{\frac{1}{r}} \quad \text{and} \quad G = \left[\prod_{k=1}^{r-1} \left(\frac{1}{\sqrt{m}} + \frac{\sqrt{m}}{k} \right)^2 \right]^{\frac{1}{r-1}}.$$

Under IEA, we have that

$$\begin{aligned} \prod_{i \leq j} Q_{ij} &= \frac{d_1(d_1 - 1) d_2(d_2 - 1) \cdots d_n(d_n - 1) 2d_1 d_2 2d_1 d_3 \cdots 2d_{n-1} d_n}{[2m(2m - 1)]^r} \\ &= \frac{2^{\binom{n}{2}} (d_1 - 1) (d_2 - 1) \cdots (d_n - 1) (d_1 d_2 \cdots d_n)^n}{[2m(2m - 1)]^r}, \end{aligned}$$

and the geometric mean is

$$\tilde{Q} = \left[\prod_{i \leq j} Q_{ij} \right]^{\frac{1}{r}} = \left[\frac{2^{\binom{n}{2}} \prod_{i=1}^n (d_i - 1) (\prod_{i=1}^n d_i)^n}{[2m(2m - 1)]^r} \right]^{\frac{1}{r}}.$$

A comparison between the distributions \mathbf{P} and \mathbf{B} for regular graphs with $n = 4$ vertices of the same degree d is shown in Figure 12 when d varies from 2 to 10 so that the number of edges $m = nd/2 = 2d$ varies from 4 to 20. There is one case where the approximate entropy

under IEA is larger than the upper bound to the exact entropy under IEA. This occurs for $m = 6$ and $\mathbf{d} = (3\ 3\ 3\ 3)$. We investigate this by using the above shown inequality. For this case we have that $r = 10$ and the geometric mean of the edge probabilities is

$$\tilde{Q} = \left[\frac{2^{\binom{n}{2}} (d-1)^n (d)^{n^2}}{[2m(2m-1)]^r} \right]^{\frac{1}{r}} = \left[\frac{2^6 2^4 3^{16}}{[12(11)]^{10}} \right]^{\frac{1}{10}} = 0.088 .$$

Further, we have that

$$G = \left[\prod_{k=1}^{r-1} \left(\frac{1}{\sqrt{m}} + \frac{\sqrt{m}}{k} \right)^2 \right]^{\frac{1}{r-1}} = \left[\prod_{k=1}^9 \left(\frac{1}{\sqrt{6}} + \frac{\sqrt{6}}{k} \right)^2 \right]^{\frac{1}{9}} = 1.107 ,$$

so that the right hand side of the inequality is equal to

$$\left(\frac{G}{2\pi e} \right)^{\frac{r-1}{r}} = \left(\frac{1.107}{2\pi e} \right)^{\frac{9}{10}} = 0.085 .$$

As seen, the geometric mean of the edge probabilities is greater than the upper bound which implies that the inequality is not satisfied, i.e. the IEA entropy approximation for this example is greater than the upper bound to the exact entropy.

Further in Figure 12, we note that as the number of edges increases, the differences between the upper bounds of the entropies and the exact or approximate entropies are increased. This indicates that the distributions of multigraphs under RSM and IEA cluster at the high probability sites when more edges are added and therefore are less flat for large values of m .

7 Simplicity and Complexity

The probability distribution of complexity of multigraphs generated by RSM depends in a complicated way on its degree sequence. Different aspects of complexity can be studied by various indicators and summary measures. For instance, the expected value of a simplicity indicator is the probability that the multigraph is simple, and it has received much attention in the literature. Janson (2009), McKay (1985), McKay and Wormald (1991), Bollobàs (1980), and Bender and Canfield (1978) all focus on asymptotic results and so far no exact solution seems to have been found. Other examples of useful information about complexity are given by the expected numbers of loops and multiple edges and their variances. This section reviews some results from the literature and presents some convenient summary measures of complexity.

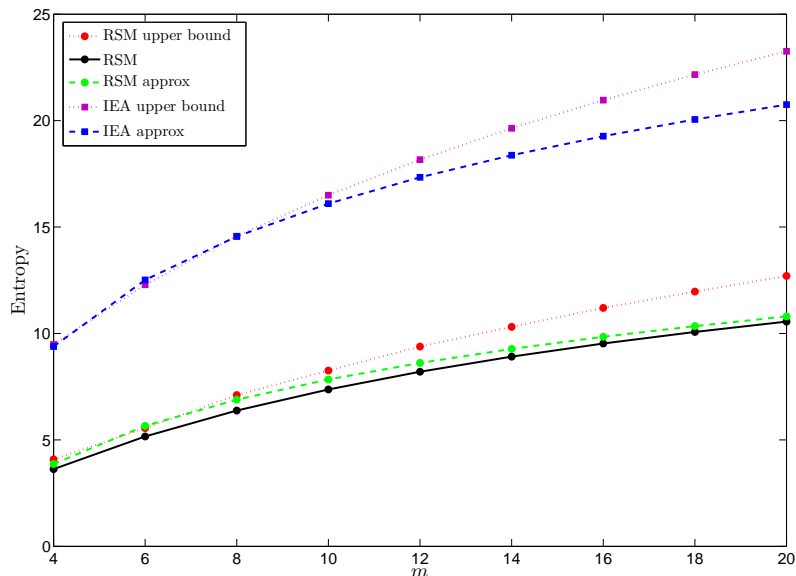


Figure 12: Approximate and exact entropies of the distribution of multigraphs under random stub matching (RSM), approximate entropies under independent edge assignments (IEA) and upper bounds of entropies in different regular multigraphs with $n = 4$ vertices. Entropies are plotted against number of edges m between 4 and 20.

There is a considerable literature about graphical degree sequences, i.e. such degree sequences that can be realized by simple graphs. Obvious necessary conditions for a finite sequence of non-negative integers (d_1, \dots, d_n) to be graphical is that $d_i < n$, $\sum_{i=1}^n d_i = 2m$ is even and $m \leq \binom{n}{2}$. Erdős and Gallai (1960) give further necessary and sufficient conditions for the existence of a simple graph with given degree sequence. They show that a degree sequence \mathbf{d} of non-negative integers in non-increasing order $d_1 \geq \dots \geq d_n$ is graphical if and only if

$$\sum_{i=1}^j d_i \leq j(j-1) + \sum_{i=j+1}^n \min(j, d_i), \quad \text{for } 1 \leq j \leq n-1.$$

Proof of necessity is straightforward; the left side of the inequality counts degree among the j vertices with highest degrees. The first term on the right side is a consequence of that at most $(j-1)$ edges are incident to any of the first j vertices. The second term of the right side is a sum of upper bounds to the number of edges for each remaining vertex. The proof of sufficiency is more complicated but can be found in several papers including Tripathi,

Venugopalan and West (2009), Sierksma and Hoogeveen (1991) and Choudum (1986).

Besides the existence result above, a recursive test to find graphical degree sequences is given by Havel (1955) and Hakimi (1962). To avoid isolated vertices only sequences with strictly positive degrees are considered in their result. They show that a non-increasing sequence $d_1 \geq \dots \geq d_n \geq 1$ with $n \geq 2$ is graphical if and only if the sequence

$$\mathbf{d}^* = (d_2 - 1, d_3 - 1, \dots, d_{d_1+1} - 1, d_{d_1+2}, \dots, d_n)$$

is graphical. The proof, which can be adapted into an algorithm to determine whether or not a sequence of positive integers can be realized by a simple graph, can be found in several papers including Blitzstein and Diaconis (2011) and Tripathi and Tyagi (2008).

The asymptotic results given by Janson (2009) concern the probability that an RSM multigraph is simple, which we denote P_0 . The asymptotic probabilities given by Janson (2009) are based on the assumptions that degrees and numbers of vertices and edges depend on some parameter that tends to infinity. The main result is that as $m \rightarrow \infty$:

$$(i) \liminf P_0 > 0 \text{ if and only if } \sum d_i^2 = O(m) ,$$

$$(ii) P_0 \rightarrow 0 \text{ if and only if } \frac{\sum d_i^2}{m} \rightarrow \infty ,$$

with the corollary that as $n \rightarrow \infty$ where $m = O(n)$ and $n = O(m)$:

$$(i) \liminf P_0 > 0 \text{ if and only if } \sum d_i^2 = O(n) ,$$

$$(ii) P_0 \rightarrow 0 \text{ if and only if } \frac{\sum d_i^2}{n} \rightarrow \infty .$$

The two asymptotic formulas for the probability that a multigraph is simple are given by Janson (2009) and they can in our notations be given as

$$P'_0 = \exp \left[- \sum_{i=1}^n \sum_{j=1}^n \lambda_{ij} + \sum_{1 \leq i < j \leq n} \log(1 + \lambda_{ij}) \right] + o(1) ,$$

and, assuming that $\max_i(d_i) = o(\sqrt{m})$,

$$P''_0 = \exp [-\Lambda(1 + \Lambda)] + o(1) ,$$

where

$$\lambda_{ij} = \frac{1}{m} \sqrt{\binom{d_i}{2} \binom{d_j}{2}}$$

and

$$\Lambda = \frac{1}{2} \sum_{i=1}^n \lambda_{ii} = \frac{1}{2m} \sum_{i=1}^n \binom{d_i}{2} = \frac{1}{4m} \sum_{i=1}^n d_i^2 - \frac{1}{2}.$$

Hence

$$P'_0 = \exp \left[-\frac{1}{2m} \left(\sum_{i=1}^n \sqrt{\binom{d_i}{2}} \right)^2 + \sum_{1 \leq i < j \leq n} \log \left(1 + \frac{1}{m} \sqrt{\binom{d_i}{2} \binom{d_j}{2}} \right) \right] + o(1),$$

and

$$P''_0 = \exp \left[-\frac{1}{4} \left(\frac{1}{2m} \sum_{i=1}^n d_i^2 \right)^2 + \frac{1}{4} \right] + o(1).$$

In particular, for regular graphs with the same degree d at every vertex, $\lambda_{ij} = (d-1)/n$ and $\Lambda = (d-1)/2$ so that

$$P'_0 = \exp \left[-\frac{n(d-1)}{2} + \binom{n}{2} \log \left(1 + \frac{(d-1)}{n} \right) \right] + o(1)$$

when $nd \rightarrow \infty$ and $n \rightarrow \infty$, and

$$P''_0 = \exp \left[-\frac{(d-1)(d+1)}{4} \right] + o(1)$$

when $d/n \rightarrow 0$ and $n \rightarrow \infty$. Some numerical examples for these approximations are presented later in this section.

Using the results obtained in Section 4 for edge multiplicities under RSM, we derive expected values and variances of some quantities that can be used to study simplicity and complexity of multigraphs. The expected values of the numbers of loops $m_1 = m_1(\boldsymbol{\eta})$ (already mentioned in previous section) and non-loops $m_2 = m_2(\boldsymbol{\eta})$ under RSM are directly obtained as expected values of local multiplicities according to:

$$E(m_1) = m \sum_{i=1}^n Q_{ii} = \frac{1}{2m-1} \sum_{i=1}^n \binom{d_i}{2}$$

and

$$E(m_2) = m \sum_{i < j} Q_{ij} = \frac{1}{2m-1} \sum_{i < j} d_i d_j.$$

We also obtain $E(m_2) = m - E(m_1)$ by using the linear relationship $m_2 = m - m_1$. This linear relationship also implies that

$$\text{Var}(m_2) = \text{Var}(m_1).$$

This common variance is given by

$$\text{Var}(m_1) = \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(m_{ii}, m_{jj}) = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^m \sum_{\ell=1}^m \text{Cov}(I_{iik}, I_{jj\ell}) ,$$

where $\text{Cov}(I_{iik}, I_{jj\ell})$ need to be considered for different i, j, k and ℓ . For $k = \ell$

$$\text{Cov}(I_{iik}, I_{jjk}) = \begin{cases} Q_{ii}(1 - Q_{ii}) & \text{for } i = j \\ -Q_{ii}Q_{jj} & \text{for } i \neq j , \end{cases}$$

and for $k \neq \ell$

$$\text{Cov}(I_{iik}, I_{jj\ell}) = \begin{cases} Q_{iiii} - Q_{ii}^2 & \text{for } i = j \\ Q_{iijj} - Q_{ii}Q_{jj} & \text{for } i \neq j , \end{cases}$$

where

$$Q_{iiii} = \frac{\binom{d_i}{2} \binom{d_i-2}{2}}{\binom{2m}{2} \binom{2m-2}{2}} \quad \text{and} \quad Q_{iijj} = \frac{\binom{d_i}{2} \binom{d_j}{2}}{\binom{2m}{2} \binom{2m-2}{2}} .$$

This implies that

$$\begin{aligned} \text{Var}(m_1) &= m \left[\sum_{i=1}^n Q_{ii}(1 - Q_{ii}) - \sum_{i \neq j} Q_{ii}Q_{jj} \right] \\ &\quad + m(m-1) \left[\sum_{i=1}^n (Q_{iiii} - Q_{ii}^2) + \sum_{i \neq j} (Q_{iijj} - Q_{ii}Q_{jj}) \right] \\ &= m \sum_{i=1}^n Q_{ii} + m(m-1) \sum_{i=1}^n Q_{iiii} - m^2 \sum_{i=1}^n Q_{ii}^2 \\ &\quad + m(m-1) \sum_{i \neq j} Q_{iijj} - m^2 \sum_{i \neq j} Q_{ii}Q_{jj} \\ &= m \sum_{i=1}^n Q_{ii}(1 - m \sum_{i=1}^n Q_{ii}) + m(m-1) \left(\sum_{i=1}^n Q_{iiii} + \sum_{i \neq j} Q_{iijj} \right) \\ &= \frac{1}{2m-1} \sum_{i=1}^n \binom{d_i}{2} \left[1 - \frac{1}{2m-1} \sum_{j=1}^n \binom{d_j}{2} \right] \\ &\quad + \frac{1}{(2m-1)(2m-3)} \left[\sum_{i=1}^n \binom{d_i}{2} \binom{d_i-2}{2} + \sum_{i \neq j} \binom{d_i}{2} \binom{d_j}{2} \right] . \end{aligned}$$

In particular, for regular graphs with the same degree d at every vertex we obtain

$$E(m_1) = \frac{d-1}{2} \left(1 + \frac{1}{nd-1} \right)$$

and

$$\text{Var}(m_1) = \frac{d-1}{2} \left(1 + \frac{1}{nd-1} + \frac{(d-2)(d-3)}{2nd} \right) + O\left(\frac{1}{n^2}\right).$$

Hence 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. The variance indicates that the number of loops might be approximately Poisson distributed.

A variable that has been used by several authors to study simplicity is $m_1 + m_3$ where m_3 is the number of pairs of equal non-loops in $\boldsymbol{\eta}$. This number is formally given by

$$m_3 = \sum_{i < j} \sum \binom{m_{ij}}{2} = \sum_{i < j} \sum_{k < \ell} I_{ijk} I_{ij\ell}.$$

The sum $m_1 + m_3$ is a variable that is 0 if and only if the multigraph is simple. Now

$$E(m_3) = \frac{m(m-1)}{2} \sum_{i < j} \sum Q_{ijij} = \frac{2}{(2m-1)(2m-3)} \sum_{i < j} \binom{d_i}{2} \binom{d_j}{2},$$

and the expected value of $m_1 + m_3$ is thus given by

$$\begin{aligned} E(m_1 + m_3) &= m \sum_{i=1}^n Q_{ii} + \binom{m}{2} \sum_{i < j} \sum Q_{ijij} \\ &= \frac{1}{2m-1} \sum_{i=1}^n \binom{d_i}{2} + \frac{2}{(2m-1)(2m-3)} \sum_{i < j} \binom{d_i}{2} \binom{d_j}{2}. \end{aligned}$$

In particular, for regular graphs with the same degree d at every vertex this expected value is about $E(m_1 + m_3) = (d^2 - 1)/4$ regardless of the number of vertices.

In order to make it easier to investigate simplicity and complexity, and find the expected value of the number $r_k = r_k(\boldsymbol{\eta})$ of sites with occupancy k , we use the IEA multiplicity distribution introduced in the previous section so that $\mathbf{m}(\boldsymbol{\eta})$ is multinomially distributed with parameters m and \mathbf{Q} . From this it follows that $m_{ij}(\boldsymbol{\eta})$ is binomially distributed with parameters m and Q_{ij} and

$$E(r_k) = \sum_{i \leq j} \sum \binom{m}{k} Q_{ij}^k (1 - Q_{ij})^{m-k} \quad \text{for } k = 0, 1, \dots, m.$$

Using this result, we obtain a formula for the expected value of the statistic $t = t(\boldsymbol{\eta})$, which in Section 3 was shown to determine the probability distribution of multigraphs under RSM. This statistic is a summary measure of complexity that is equal to 0 if and only if the multigraph is simple. Its expected value under IEA is given by

$$\begin{aligned} E(t) &= E \left[m_1 + \sum_{k=2}^m r_k \log k! \right] \\ &= m \sum_{i=1}^n Q_{ii} + \sum_{k=2}^m \left[\log k! \binom{m}{k} \sum_{i \leq j} Q_{ij}^k (1 - Q_{ij})^{m-k} \right]. \end{aligned}$$

Note that the expected value of t under RSM that was given in the previous section is considerably more complicated to specify and analyze.

Other statistics related to complexity are also easily handled under IEA. The number of sites with no occupancy given by $r_0 = r_0(\boldsymbol{\eta})$ has expected value

$$E(r_0) = \sum_{i \leq j} (1 - Q_{ij})^m,$$

and the number of sites with single occupancy given by $r_1 = r_1(\boldsymbol{\eta})$ has expected value

$$E(r_1) = m \sum_{i \leq j} Q_{ij} (1 - Q_{ij})^{m-1}.$$

The expected value of the number of multiple occupancy sites is thus

$$E(r - r_0 - r_1) = r - \sum_{i \leq j} (1 - Q_{ij})^m - m \sum_{i \leq j} Q_{ij} (1 - Q_{ij})^{m-1},$$

and the number of multiple edges has expected value

$$E(m - r_1) = m \left(1 - \sum_{i \leq j} Q_{ij} (1 - Q_{ij})^{m-1} \right).$$

A particularly interesting statistic is r_{21} (defined in Section 3) which is equal to m if and only if the multigraph is simple. This means that there are m single occupancies of non-loops. The exact probability distribution of this statistic is unknown, but being a counting statistic it is not unreasonable to assume that it is approximately Poisson distributed with parameter $\lambda = E(r_{21})$. The probability of simplicity $P_0 = P(r_{21} = m)$ can then be approximated by

$$P_0''' = \frac{e^{-\lambda} \lambda^m}{m!},$$

where

$$\lambda = E(r_{21}) = m \sum_{i \leq j} Q_{ij} (1 - Q_{ij})^{m-1} = m \sum_{i < j} \frac{d_i d_j}{m(2m-1)} \left(1 - \frac{d_i d_j}{m(2m-1)}\right)^{m-1}.$$

For regular graphs with the same degree d at every vertex we obtain

$$E(r_{21}) = \frac{\binom{n}{2} d^2}{(nd-1)} \left(1 - \frac{2d}{n(nd-1)}\right)^{\frac{nd}{2}-1}.$$

This expected value is for large n and small d approximately equal to

$$E(r_{21}) \approx \frac{\binom{n}{2} d^2}{(nd-1)} \exp\left(-\frac{d(nd-2)}{n(nd-1)}\right) \approx \frac{(n-1)d}{2} \exp\left(-\frac{d}{n}\right) \approx \frac{nd}{2} - \frac{d(d+1)}{2} + \frac{d^2(d+2)}{4n}$$

and it follows for this case that

$$P_0''' = \frac{e^\lambda \lambda^m}{m!} \approx \frac{\lambda^m e^{m-\lambda}}{m^m \sqrt{2\pi m}} \approx \frac{e^{\binom{d+1}{2}}}{\sqrt{2\pi m}} \left(1 - \frac{\binom{d+1}{2}}{m}\right)^m.$$

In Table 9 we give some numerical examples of the probability that a RSM multigraph is simple. These probabilities are compared to the previously given asymptotic probabilities P_0' and P_0'' and the approximate probability P_0''' . Here, we look at small graphs with 6 to 8 vertices, and study cases where the numbers of edges are in the interval ± 1 of the number of vertices. For each case presented in Table 9, we focus on degree sequences with positive degrees at each vertex that are at most 3 so that a reasonable number of simple graphs is possible. From Table 9 we see that the Poisson approximation is close to the RSM probability of simplicity but, as expected, the asymptotic probabilities do not perform well for these small examples. In particular, the best Poisson approximations are for cases where $m = n + 1$. Further we note that the Poisson approximations do not perform well for the regular graphs presented in Table 9.

Table 9: Some numerical examples of the probability that an RSM multigraph is simple, compared to the suggested Poisson approximation of m single edge occupancies of non-loops, and the two asymptotic probabilities suggested by Janson (2009).

$n = 6, m = 5$				
Degree sequence	RSM	Poisson	Asymptotic 1	Asymptotic 2
$\mathbf{d} = (3, 3, 1, 1, 1, 1)$	0.107	0.104	0.482	0.383
$\mathbf{d} = (3, 2, 2, 1, 1, 1)$	0.195	0.116	0.540	0.472
$\mathbf{d} = (2, 2, 2, 2, 1, 1)$	0.284	0.128	0.603	0.571
$n = 6, m = 6$				
Degree sequence	RSM	Poisson	Asymptotic 1	Asymptotic 2
$\mathbf{d} = (3, 3, 3, 1, 1, 1)$	0.042	0.070	0.356	0.269
$\mathbf{d} = (3, 3, 2, 2, 1, 1)$	0.086	0.082	0.401	0.329
$\mathbf{d} = (3, 2, 2, 2, 2, 1)$	0.132	0.093	0.450	0.397
$\mathbf{d} = (2, 2, 2, 2, 2, 2)$	0.180	0.104	0.503	0.472
$n = 6, m = 7$				
Degree sequence	RSM	Poisson	Asymptotic 1	Asymptotic 2
$\mathbf{d} = (3, 3, 3, 3, 1, 1)$	0.031	0.051	0.276	0.204
$\mathbf{d} = (3, 3, 3, 2, 2, 1)$	0.047	0.061	0.311	0.246
$\mathbf{d} = (3, 3, 2, 2, 2, 2)$	0.069	0.070	0.349	0.294
$n = 7, m = 6$				
Degree sequence	RSM	Poisson	Asymptotic 1	Asymptotic 2
$\mathbf{d} = (3, 3, 2, 1, 1, 1, 1)$	0.132	0.100	0.473	0.397
$\mathbf{d} = (3, 2, 2, 2, 1, 1, 1)$	0.200	0.110	0.526	0.472
$\mathbf{d} = (2, 2, 2, 2, 2, 1, 1)$	0.270	0.119	0.582	0.554
$n = 7, m = 7$				
Degree sequence	RSM	Poisson	Asymptotic 1	Asymptotic 2
$\mathbf{d} = (3, 3, 3, 2, 1, 1, 1)$	0.068	0.076	0.365	0.294
$\mathbf{d} = (3, 3, 2, 2, 2, 1, 1)$	0.103	0.085	0.406	0.348
$\mathbf{d} = (3, 2, 2, 2, 2, 2, 1)$	0.143	0.093	0.451	0.407
$\mathbf{d} = (2, 2, 2, 2, 2, 2, 2)$	0.189	0.102	0.499	0.472
$n = 7, m = 8$				
Degree sequence	RSM	Poisson	Asymptotic 1	Asymptotic 2
$\mathbf{d} = (3, 3, 3, 3, 2, 1, 1)$	0.043	0.059	0.291	0.229
$\mathbf{d} = (3, 3, 3, 2, 2, 2, 1)$	0.060	0.067	0.324	0.269
$\mathbf{d} = (3, 3, 2, 2, 2, 2, 2)$	0.082	0.075	0.360	0.313
$n = 8, m = 7$				
Degree sequence	RSM	Poisson	Asymptotic 1	Asymptotic 2
$\mathbf{d} = (3, 3, 3, 1, 1, 1, 1, 1)$	0.098	0.090	0.424	0.348
$\mathbf{d} = (3, 3, 2, 2, 1, 1, 1, 1)$	0.148	0.098	0.469	0.407
$\mathbf{d} = (3, 2, 2, 2, 2, 1, 1, 1)$	0.202	0.106	0.516	0.472
$\mathbf{d} = (2, 2, 2, 2, 2, 2, 1, 1)$	0.262	0.113	0.566	0.542
$n = 8, m = 8$				
Degree sequence	RSM	Poisson	Asymptotic 1	Asymptotic 2
$\mathbf{d} = (3, 3, 3, 3, 1, 1, 1, 1)$	0.059	0.071	0.337	0.269
$\mathbf{d} = (3, 3, 3, 2, 2, 1, 1, 1)$	0.084	0.079	0.373	0.313
$\mathbf{d} = (3, 3, 2, 2, 2, 2, 1, 1)$	0.115	0.086	0.411	0.362
$\mathbf{d} = (3, 2, 2, 2, 2, 2, 2, 1)$	0.151	0.093	0.452	0.415
$\mathbf{d} = (2, 2, 2, 2, 2, 2, 2, 2)$	0.193	0.099	0.496	0.472
$n = 8, m = 9$				
Degree sequence	RSM	Poisson	Asymptotic 1	Asymptotic 2
$\mathbf{d} = (3, 3, 3, 3, 3, 1, 1, 1)$	0.040	0.058	0.275	0.217
$\mathbf{d} = (3, 3, 3, 3, 2, 2, 1, 1)$	0.053	0.065	0.305	0.251
$\mathbf{d} = (3, 3, 3, 2, 2, 2, 2, 1)$	0.071	0.071	0.336	0.288
$\mathbf{d} = (3, 3, 2, 2, 2, 2, 2, 2)$	0.093	0.078	0.369	0.329

A convenient way to obtain the IEA multiplicity distribution is to assume that the stubs are randomly generated and independently assigned to vertices, independent stub assignments (ISA). If stubs ξ_k for $k = 1, \dots, 2m$ are independently and identically distributed according to a probability distribution $\mathbf{p} = (p_1, \dots, p_n)$ with positive probabilities for the n vertices, it follows that the sequence of stub frequencies $\mathbf{d}(\boldsymbol{\xi})$ is multinomially distributed with parameters $2m$ and \mathbf{p} . It also follows that edges in $\boldsymbol{\xi}$ are independent and equal to (i, j) with probabilities $P_{ij} = p_i p_j$ for $i = 1, \dots, n$ and $j = 1, \dots, n$. Edges in $\boldsymbol{\eta}$ are independent and equal to (i, j) with probabilities $Q_{ij} = p_i^2$ for $i = j$, and $Q_{ij} = 2p_i p_j$ for $i < j$. Now the edge multiplicity sequence $\mathbf{m}(\boldsymbol{\eta})$ is multinomially distributed with parameters m and \mathbf{Q} . This is an IEA distribution with a new \mathbf{Q} based on ISA. The conditional distribution of $\mathbf{m}(\boldsymbol{\eta})$ given $\mathbf{d}(\boldsymbol{\xi})$ is equal to the previous edge multiplicity distribution obtained by random stub matching with fixed degrees. This is a consequence of that the conditional probabilities under ISA and IEA can be transformed according to

$$P(\mathbf{m}(\boldsymbol{\eta}) = \mathbf{m} | \mathbf{d}(\boldsymbol{\xi}) = \mathbf{d}) = \frac{\binom{m}{\mathbf{m}} \mathbf{Q}^{\mathbf{m}}}{\binom{2m}{\mathbf{d}} \mathbf{p}^{\mathbf{d}}} = \frac{\binom{m}{\mathbf{m}} 2^{m_2}}{\binom{2m}{\mathbf{d}}} ,$$

using that

$$\mathbf{Q}^{\mathbf{m}} = \prod_{i \leq j} Q_{ij}^{m_{ij}} = \left(\prod_{i=1}^n p_i^{2m_{ii}} \right) \left(\prod_{i < j} (2p_i p_j)^{m_{ij}} \right) = 2^{m_2} \prod_{i=1}^n p_i^{d_i} = 2^{m_2} \mathbf{p}^{\mathbf{d}}$$

and $m_2 = \sum \sum_{i < j} m_{ij}$. The multinomial distribution for $\mathbf{d}(\boldsymbol{\xi})$ with parameters $2m$ and \mathbf{p} can be considered as a Bayesian model for the stub frequencies.

By using that the sequence of stub frequencies $\mathbf{d}(\boldsymbol{\xi})$ is multinomially distributed with parameters $2m$ and \mathbf{p} under ISA, we derive a formula for the expected entropy of the distribution of multigraphs $\mathbf{m}(\boldsymbol{\eta})$. This expected value is found by using the expected entropy under ISA which is equal to the difference $H(\mathbf{m}) - H(\mathbf{d})$ using calculation rules for entropy (given for instance in Frank 2011). Under ISA we use normal approximations to the multinomial distributions of \mathbf{m} and \mathbf{d} and obtain the approximate entropies

$$H(\mathbf{m}) \approx \log \sqrt{(2\pi e m)^{r-1} \prod_{i \leq j} Q_{ij}}$$

and

$$H(\mathbf{d}) \approx \log \sqrt{(4\pi e m)^{n-1} \prod_i p_i} .$$

It follows that the expected entropy under ISA is given by

$$\begin{aligned}
E [H(\mathbf{m}|\mathbf{d})] &\approx \log \sqrt{(2\pi em)^{r-1} \prod_{i \leq j} Q_{ij}} - \log \sqrt{(4\pi em)^{n-1} \prod_{i=1}^n p_i} \\
&= \log \sqrt{\frac{(2\pi em)^{r-n} \prod_{i \leq j} Q_{ij}}{2^{n-1} \prod_{i=1}^n p_i}} \\
&= \log \sqrt{\frac{(2\pi em)^{\binom{n}{2}} 2^{\binom{n}{2}} (p_1 \cdots p_n)^{n+1}}{2^{n-1} (p_1 \cdots p_n)}} \\
&= \log \sqrt{(2\pi em)^{\binom{n}{2}} 2^{\binom{n-1}{2}} (p_1 \cdots p_n)^n}.
\end{aligned}$$

For fixed n and \mathbf{p} this is a linear expression in $\log m$ which is denoted

$$H^* = a^* + b^* \log m ,$$

where

$$a^* = a^*(n, \mathbf{p}) = \log \sqrt{(2\pi e)^{\binom{n}{2}} 2^{\binom{n-1}{2}} (p_1 \cdots p_n)^n}$$

and

$$b^* = b^*(n) = n(n-1)/4 .$$

If this expected entropy is calculated with $\mathbf{p} = \mathbf{d}/2m$ for a fixed degree sequence \mathbf{d} , it can be considered as an approximation to the entropy H of the distribution of multigraphs \mathbf{m} under RSM with this degree sequence \mathbf{d} . In particular, for the distribution of multigraphs under RSM with all $d_i = 2m/n$, its entropy is approximately given by

$$H^* = \log \sqrt{(2\pi em)^{\binom{n}{2}} 2^{\binom{n-1}{2}} n^{-n^2}} ,$$

corresponding to uniform \mathbf{p} .

Another approximation H^{**} to the entropy H of the distribution of multigraphs \mathbf{m} under RSM can be based on asymptotic results for the expected entropy, $E [H(\mathbf{m}|\mathbf{d})] = H(\mathbf{m}) - H(\mathbf{d})$, under ISA. Both \mathbf{m} and \mathbf{d} are multinomially distributed and obtained from sequences of m independent identically distributed sites, and $2m$ independent identically distributed stubs. The central limit theorem in weak form yields asymptotic equipartition properties for r_c and n_c stubs. Here r_c and n_c are given by the entropies of site and stub probabilities according to

$$\log r_c = h(\mathbf{Q}) \quad \text{and} \quad \log n_c = h(\mathbf{p}) .$$

In particular, for uniform ISA it follows that $n_c = n$ and $r_c = n^2 2^{-(n-1)/n}$ which is about $\binom{n+1}{2}$ for large n . See for instance Cover and Thomas (1991) which has a detailed presentation of the equipartition property. The equipartition property implies that \mathbf{m} and \mathbf{d} are

asymptotically multinomially distributed with r_c equal site probabilities and n_c equal stub probabilities. Under ISA it holds that

$$h(\mathbf{Q}) = 2h(\mathbf{p}) - 1 + \sum_{i=1}^n p_i^2 ,$$

and hence

$$r_c = n_c^2 2^{-(n_c-1)/n_c} .$$

The asymptotic entropies of \mathbf{m} and \mathbf{d} under ISA are thus given by

$$H_c(\mathbf{m}) = \log \sqrt{(2\pi em)^{r_c-1} r_c^{-r_c}}$$

and

$$H_c(\mathbf{d}) = \log \sqrt{(4\pi em)^{n_c-1} n_c^{-n_c}} .$$

The difference

$$H_c(\mathbf{m}) - H_c(\mathbf{d}) = \log \sqrt{\frac{(2\pi em)^{r_c-1} n_c^{n_c}}{(4\pi em)^{n_c-1} r_c^{r_c}}}$$

is the asymptotic expected conditional entropy of \mathbf{m} given \mathbf{d} , when \mathbf{d} is obtained from \mathbf{p} . The RSM entropy H of \mathbf{m} with a fixed \mathbf{d} can be approximated by the asymptotic expression obtained with $\mathbf{p} = \mathbf{d}/2m$ for this fixed \mathbf{d} . Let H^{**} denote this approximation. It can be given as

$$H^{**} = a^{**} + b^{**} \log m ,$$

where

$$a^{**} = a^{**}(n, h(\mathbf{p})) = \log \sqrt{\frac{(2\pi e)^{r_c-1} n_c^{n_c}}{(4\pi e)^{n_c-1} r_c^{r_c}}} ,$$

and

$$b^{**} = b^{**}(n, h(\mathbf{p})) = \frac{r_c - n_c}{2} .$$

Now $r_c = n^2 2^{-(n_c-1)/n_c}$ where $n_c = 2^{h(\mathbf{p})}$ and \mathbf{p} is chosen as $\mathbf{p} = \mathbf{d}/2m$ in order to estimate the RSM entropy H with this \mathbf{d} . Note that the dependency of H^{**} on \mathbf{p} is only via the degree distribution entropy $h(\mathbf{p})$.

Using these results, a few examples are given to illustrate the performance of the approximations H^* and H^{**} to the entropy H of the distribution of multigraphs under RSM. In Table 10 we compare two approximations using multigraphs with $n = 3$ vertices and $m = 6$ edges. Thus, the approximations for this case are given by

$$H^* = \log \sqrt{(\pi e)^3 2 (d_1 \cdot d_2 \cdot d_3)^3 (12)^{-6}}$$

and

$$H^{**} = \log \sqrt{\frac{(2\pi e 6)^{r_c-1} n_c^{n_c}}{(4\pi e 6)^{n_c-1} r_c^{r_c}}},$$

where $\mathbf{d} = (d_1, d_2, d_3) = 2m\mathbf{p}$ is varied. As seen in Table 10, both approximations are good for degree distributions that are uniformly or nearly uniformly distributed. Note that the approximation H^* is not useful for very skew degree distributions, e.g. $\mathbf{d} = (10, 1, 1)$, which yields a negative entropy approximation. However, H^{**} yields better approximations for these skew degree distributions.

Table 10: The entropy of the distribution of multigraphs under random stub matching (RSM) and its approximations in multigraphs with $n = 3$ vertices and $m = 6$ edges for various degree sequences \mathbf{d} .

$\mathbf{d} = 2m\mathbf{p}$	Entropy	Expected entropy	Asymptotic entropy
	RSM (\mathbf{d})	ISA (\mathbf{p})	ISA (\mathbf{p})
	H	H^*	H^{**}
(10, 1, 1)	0.440	-0.631	0.754
(9, 2, 1)	1.096	0.641	1.247
(8, 3, 1)	1.651	1.264	1.665
(7, 4, 1)	2.044	1.597	1.965
(6, 5, 1)	2.242	1.747	2.121
(7, 3, 2)	2.362	2.475	2.343
(6, 4, 2)	2.723	2.763	2.642
(5, 5, 2)	2.847	2.852	2.744
(6, 3, 3)	2.889	3.019	2.815
(5, 4, 3)	3.172	3.247	3.057
(4, 4, 4)	3.330	3.386	3.197

To visualize these findings further, we conclude this section with comparisons between the entropy under RSM and the approximations. First, consider regular multigraphs with $n = 4$ vertices having the same degree d that varies from 1 to 15, so that the number of edges $m = nd/2 = 2d$ varies from 2 to 30. Thus, the degree sequences of these multigraphs are uniformly distributed under ISA with $p_i = 1/n$. In Figure 13 where we see that the entropy under RSM is well approximated by both $H^* = a^* + b^* \log m = -2.22 + 3 \log m$, and $H^{**} = a^{**} + b^{**} \log m = -1.67 + 2.76 \log m$. The RSM entropy deviates slightly from linearity in $\log m$, which is easier to see in Figure 14 where the entropy and its approximations are plotted against $\log m$.

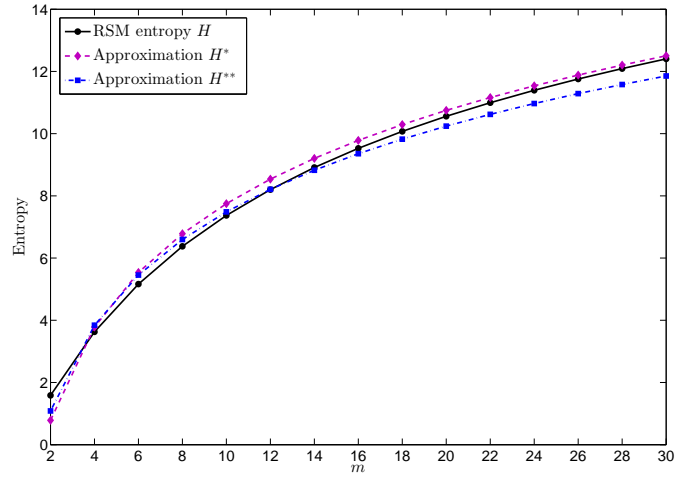


Figure 13: The entropy of the distribution of multigraphs under random stub matching (RSM) and its approximations for different regular multigraphs with $n = 4$ vertices and different numbers of edges m between 2 and 30.

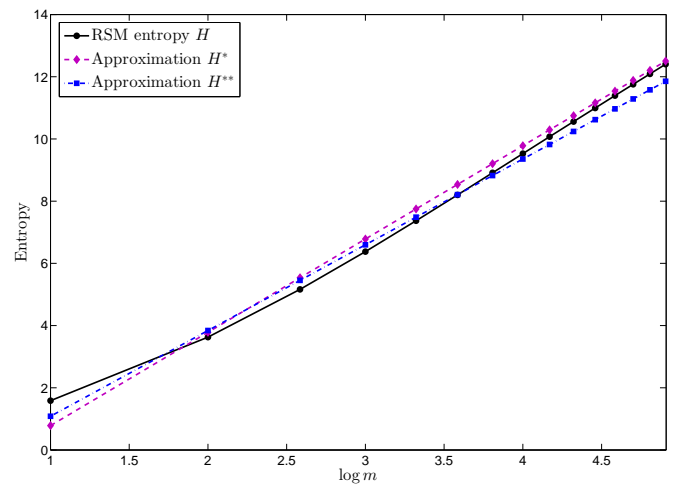


Figure 14: The entropy of the distribution of multigraphs under random stub matching (RSM) and its approximations for different regular multigraphs with $n = 4$ vertices and different numbers of edges m between 2 and 30. Entropy is plotted against $\log m$.

In Figure 15 and 16, we consider two cases with skew degree distributions in multigraphs with $n = 3$ and $n = 4$ vertices. Here, $\mathbf{p} = (4/7, 2/7, 1/7)$ and $\mathbf{p} = (5/10, 3/10, 1/10, 1/10)$ which implies that possible degree sequences for RSM are multiples of the degree sequences $\mathbf{d} = (8, 4, 2)$ and $\mathbf{d} = (5, 3, 1, 1)$, respectively. As seen from these two figures, the RSM entropies are well approximated by H^* , but not by H^{**} which deviates more and more from RSM entropy as m increases. When m increases, the same \mathbf{p} will result in different degree sequences \mathbf{d} which in turn give entropies that are harder to approximate by the asymptotic method. From these examples we also note that both the approximations can be either larger or smaller than the RSM entropy H .

In Table 11 we compare the RSM entropy and its approximations using multigraphs with $n = 8$ vertices and $m = 8$ edges. It is clear that approximations H^{**} perform much better than approximations H^* , in particular for the skew degree distributions. However, as the degree sequences become more uniformly distributed, the performance of H^* is improved.

The findings from these considered cases can be summarized as follows. The approximation H^* performs well for multigraphs with degree distributions that are uniformly or close to uniformly distributed. This holds for all multigraphs, no matter size. For skew degree distributions, H^* performs well if the edge frequency m is large. The approximation H^{**} performs well for small multigraphs, i.e. multigraphs with small numbers of vertices and edges, no matter degree distributions. Further, if the number of vertices n increases, this approximation is much better than H^* . If the degree distribution is uniformly distributed, H^{**} also performs well for small number of vertices but large edge frequencies.

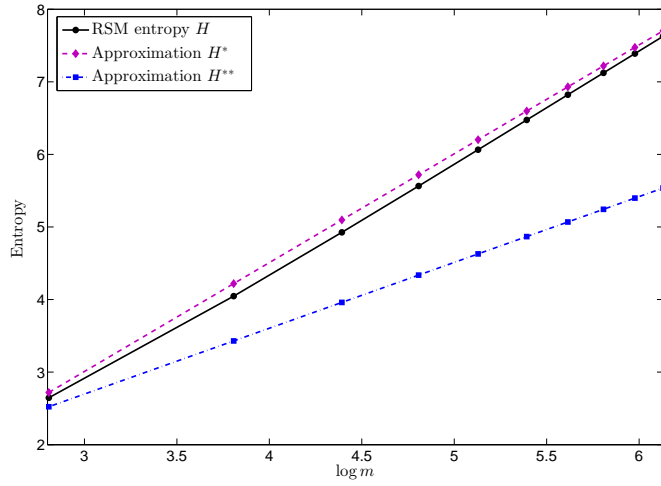


Figure 15: The entropy of the distribution of multigraphs under random stub matching (RSM) and its approximations for multigraphs with $n = 3$ vertices and degree sequences that are multiples of $\mathbf{d} = (8, 4, 2)$ for edge frequencies $m = 7, 14, 21, \dots, 70$. Entropy is plotted against $\log m$.

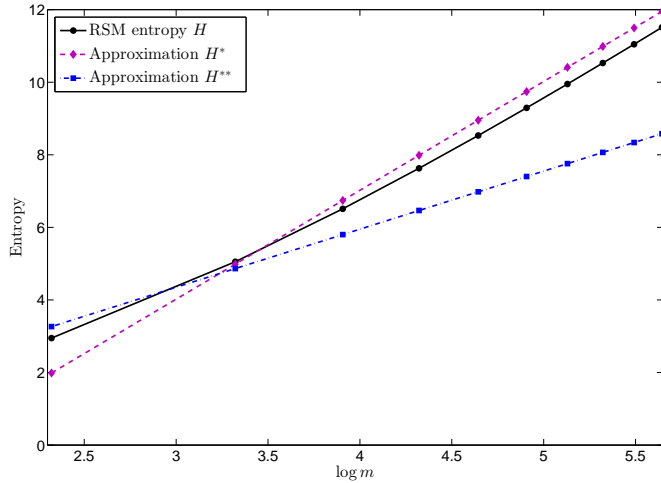


Figure 16: The entropy of the distribution of multigraphs under random stub matching (RSM) and its approximations for multigraphs with $n = 4$ vertices and degree sequences that are multiples of $\mathbf{d} = (5, 3, 1, 1)$ for edge frequencies $m = 5, 10, 15, \dots, 50$. Entropy is plotted against $\log m$.

Table 11: The entropy of the distribution of multigraphs under random stub matching (RSM) and its approximations in multigraphs with $n = 8$ vertices and $m = 8$ edges for various degree sequences \mathbf{d} .

$\mathbf{d} = 2m\mathbf{p}$	Entropy	Expected entropy	Asymptotic entropy
	RSM (\mathbf{d})	ISA (\mathbf{p})	ISA (\mathbf{p})
	H	H^*	H^{**}
(9, 1, 1, 1, 1, 1, 1, 1)	6.589	-5.502	8.447
(8, 2, 1, 1, 1, 1, 1, 1)	7.869	-2.181	9.897
(7, 3, 1, 1, 1, 1, 1, 1)	8.790	-0.612	10.82
(6, 4, 1, 1, 1, 1, 1, 1)	9.352	0.1589	11.340
(5, 5, 1, 1, 1, 1, 1, 1)	9.541	0.394	11.505
(7, 2, 2, 1, 1, 1, 1, 1)	9.144	1.048	11.259
(6, 3, 2, 1, 1, 1, 1, 1)	9.992	2.498	12.037
(5, 4, 2, 1, 1, 1, 1, 1)	10.419	3.106	12.391
(5, 3, 3, 1, 1, 1, 1, 1)	10.701	3.786	12.646
(4, 4, 3, 1, 1, 1, 1, 1)	10.938	4.159	12.831
(6, 2, 2, 2, 1, 1, 1, 1)	10.379	4.159	12.443
(5, 3, 2, 2, 1, 1, 1, 1)	11.109	5.446	13.023
(4, 4, 2, 2, 1, 1, 1, 1)	11.352	5.819	13.195
(4, 3, 3, 2, 1, 1, 1, 1)	11.652	6.498	13.417
(3, 3, 3, 3, 1, 1, 1, 1)	11.958	7.178	13.625
(5, 2, 2, 2, 2, 1, 1, 1)	11.526	7.106	13.373
(4, 3, 2, 2, 2, 1, 1, 1)	12.084	8.159	13.731
(3, 3, 3, 2, 2, 1, 1, 1)	12.398	8.838	13.914
(4, 2, 2, 2, 2, 2, 1, 1)	12.525	9.819	14.005
(3, 3, 2, 2, 2, 2, 1, 1)	12.847	10.498	14.159
(3, 2, 2, 2, 2, 2, 2, 1)	13.304	12.159	14.351
(2, 2, 2, 2, 2, 2, 2, 2)	13.771	13.819	14.482

References

- Bayati, M., Kim, J.H. and Saberi, A. (2010), A Sequential Algorithm for Generating Random Graphs, *Algorithmica*, **58**, 860–910.
- Bender, E. A. and Canfield, E. R. (1978), The Asymptotic Number of Labeled Graphs with Given Degree Sequences, *Journal of Combinatorial Theory Series A*, **24(3)**, 296–307.
- Blitzstein, J. and Diaconis, P. (2011), A Sequential Importance Sampling Algorithm for Generating Random Graphs with Prescribed Degrees, *Internet Mathematics*, **6(4)**, 489–522.
- Bollobàs, B. (1980), A Probabilistic Proof of an Asymptotic Formula for the Number of Labelled Regular Graphs, *European Journal of Combinatorics*, **1(4)**, 311–316.
- Bollobàs, B. (2001), *Random Graphs*, Second Edition, Cambridge: Cambridge University Press.
- Britton, T., Deijfen, M. and Martin-Löf A. (2006), Generating Simple Random Graphs with Prescribed Degree Distribution, *Journal of Statistical Physics*, **124(6)**, 1377–1397.
- Choudum, S. A. (1986), A Simple Proof of the Erdős-Gallai Theorem on Graph Sequences, *Bulletin of the Australian Mathematical Society*, **33(1)**, 67–70.
- Chung, F. and Lu, L. (2002), Connected Components in Random Graphs with Given Expected Degree Sequences, *Annals of Combinatorics*, **6**, 125–145.
- Cover, T. and Thomas, J. (1991), *Elements of Information Theory*, New York: Wiley Series in Communication.
- Erdős, P. and Gallai, T. (1960), Graphen mit Punkten Vorgeschiedenen Grades, *Matematikai Lapok*, **11**, 264–274.
- Frank, O. (2011), Statistical Information Tools for Multivariate Discrete Data, in *Modern Mathematical Tools and Techniques in Capturing Complexity*, eds. L. Pardo, N. Balakrishnan and M. Ángeles Gil, Berlin: Springer Verlag, 177–190.
- Frank, O. and Nowicki, K. (1989), On Entropies of Occupancy Distributions, in *Combinatorics and Graph Theory*, eds. Zdzislaw Skupien, Mieczyslaw Borowiecki, Banach Center Publications, **25**, PWN-Polish Scientific Publishers, Warsaw.
- Frank, O. and Shafie, T. (2012), Complexity of Families of Multigraphs, to appear in *JSM Proceedings*, Section on Statistical Graphics, Alexandria, VA: American Statistical Association.
- Hakimi, S. L. (1962), On Realizability of a Set of Integers as Degrees of the Vertices of a Linear Graph I, *Journal of the Society for Industrial and Applied Mathematics*, **10(3)**, 496–506.
- Havel, V. (1955), A Remark on the Existence of Finite Graphs, *Casopis Pest. Mat.*, **80**, 477–480.
- Janson, S. (2009), The Probability that a Random Multigraph is Simple, *Combinatorics, Probability and Computing*, **18(1–2)**, 205–225.
- McKay B. D. (1985), Asymptotics for Symmetric 0-1 Matrices with Prescribed Row Sums, *Ars Combinatoria*, **19A**, 15–25.

McKay, B. D. and Wormald, N. C. (1991), Asymptotic Enumeration by Degree Sequence of Graphs with degrees $o(n^{1/2})$, *Combinatorica*, **11(4)**, 369–382.

Sierksma, G. and Hoogeveen, H. (1991), Seven Criteria for Integer Sequences being Graphic, *Journal of Graph Theory*, **15(2)**, 223–231.

Tripathi, A. and Tyagi, H. (2008), A Simple Criterion on Degree Sequences of Graphs, *Discrete Applied Mathematics*, **156(18)**, 3513–3517.

Tripathi, A., Venugopalanb, A. and West, D. B. (2010), A Short constructive proof of the Erdős-Gallai Characterization of Graphic Lists, *Discrete Mathematics*, **310(4)**, 833–834.

Statistical Analysis of Multigraphs

Termeh Shafie

Abstract

This article analyzes multigraphs by performing statistical tests of multigraph models obtained by random stub matching (RSM) and by independent edge assignments (IEA). The tests are performed using goodness-of-fit measures between the multiplicity sequence of an observed multigraph and the expected multiplicity sequence according to a simple or composite IEA hypothesis. Test statistics of Pearson type and of information divergence type are used. The expected values of the Pearson goodness-of-fit statistic under different multigraph models are derived, and some approximations of the test statistics with adjusted χ^2 -distributions are considered. Illustrations of test performances are presented for all models, and the results indicate that even for very small number of edges, the null distributions of both statistics are well approximated by their asymptotic χ^2 -distribution. This holds true for testing simple as well as composite hypotheses with different asymptotic distributions. The non-null distributions of the test statistics can be well approximated by adjusted χ^2 -distributions which can be used for power approximations. The influence of RSM on both test statistics is substantial for small number of edges and implies a shift of their distributions towards smaller values compared to what holds true for the null distributions under IEA.

Keywords: multigraph, multiplicity, goodness-of-fit, information divergence.

1 Introduction

A random multigraph model is given by a probability distribution over some class of multigraphs. In this article multigraphs are analyzed by performing statistical tests of some multigraph models presented in Frank and Shafie (2012) and Shafie (2012). Two main multigraph models are considered. The first is obtained by random stub matching with fixed degrees (RSM) so that edge assignments to sites are dependent, and the second is obtained by independent edge assignments (IEA) according to a common probability distribution. Further, we present two different methods for obtaining an approximate IEA model from an RSM model. This is done by assuming that the stubs are randomly generated

Department of Statistics, Stockholm University, S-106 91 Stockholm, termeh.shafie@stat.su.se

and independently assigned to vertices (ISA) and can be viewed as a Bayesian model for the stub frequencies under RSM. Another way of obtaining an approximate IEA model is to ignore the dependency between edges in the RSM model and assume independent edge assignments of stubs (IEAS). The tests are performed using goodness-of-fit measures between the multiplicity sequence of an observed multigraph and the expected multiplicity sequence according to a simple or composite IEA hypothesis. The exact distributions of the test statistics are investigated and compared to different approximations given by adjusted χ^2 -distributions.

In the next section, multigraph data structures are described and exemplified. It is shown how they can be obtained by different kinds of vertex and edge aggregations. These kinds of aggregations are powerful methods to analyze structures in very large graphs. In Section 3, some basic notations are introduced, and the different multigraph models mentioned above are defined.

Statistical tests of simple hypotheses are considered in Section 4 where the hypotheses are fully specified IEA models. For an IEAS model, the edge probability parameters are functions of a specified degree sequence \mathbf{d} , and for an ISA model these parameters are functions of a specified stub selection probability sequence \mathbf{p} . The Pearson goodness-of-fit statistic S and the divergence statistic T for these tests are defined. The expected value of S is derived under different multigraph models, and in particular it is shown that for the null distribution under RSM, this expected value only depends on the numbers of vertices and edges. Test illustrations for IEAS, ISA and RSM models are presented where the moments and cumulative distribution functions of the test statistics are used to compare and evaluate their performances. The convergence of the null distributions of S and T to their asymptotic χ^2 -distributions is rapid and even for small number of edges m , a good fit is seen between the null distributions and the asymptotic χ^2 -distribution. For cases when flat \mathbf{d} or \mathbf{p} is tested against skew \mathbf{d} or \mathbf{p} (or vice versa), both statistics have good powers of rejecting a simple hypothesis about a false model. The non-null distributions of S and T needed for determining power are approximated by adjusted χ^2 -distributions. The influence of RSM on the distributions of S and T is substantial for small m and implies a shift towards smaller values of the statistics compared to what holds true for the null distributions under IEA.

In Section 5, statistical tests of composite multigraph hypotheses are illustrated for IEAS, ISA and RSM models. Moments and cumulative distribution functions of the test statistics are used for comparisons and evaluations of their performances. The composite multigraph hypotheses might be unspecified IEAS or ISA where the parameters have to be estimated from data. For composite IEAS or ISA hypotheses including the correct model, the following results are noted. The null distributions of S and T converge faster to their asymptotic χ^2 -distributions for flat \mathbf{d} or \mathbf{p} than for skew \mathbf{d} or \mathbf{p} , but even for rather small m , there is a good fit between these distributions and their asymptotic χ^2 -distributions. Further, both statistics have very poor powers of detecting differences between IEAS and ISA hypotheses for small as well as for large m .

2 Data Structures and Possible Applications

A multigraph is defined as a graph where multiple edges and edge-loops are permitted. Such data structures are common in contexts when several edges can be mapped on the same vertex pair, but they are also obtained by different types of aggregation. Several simple graphs representing different binary relations can be aggregated to a multigraph, or an initial very large graph can be transformed to a multigraph by aggregating vertices into special subsets. Such possibilities are illustrated by some examples.

Consider a social network of friendships between 15 school children consisting of 12 pairs of mutual friendships. The children are categorized by two attributes, gender with categories labeled G (girl) and B (boy), and living area with categories labeled N (north) and S (south). Thus, there are four vertex categories BN, BS, GN and GS which are displayed together with mutual friendships in Figure 1. By aggregating vertices in the same category, we obtain a multigraph on 4 new vertices corresponding to the categories, and it has the same number of edges as the initial graph. This is shown in Figure 2. By performing this kind of transformation, we reduce the number of vertices but increase the number of multiple edges and edge loops. Generally, social networks of contacts between individuals can be transformed to multigraphs on vertices corresponding to combined categories of individual attributes, and edge multiplicities represent frequencies of contacts within and between these categories.

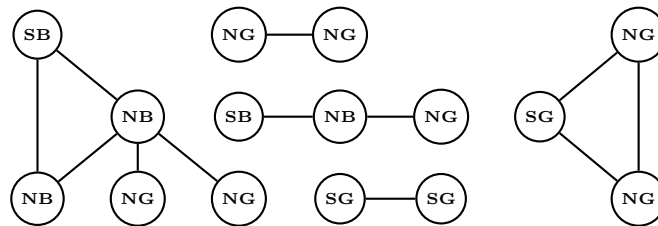


Figure 1: Initial graph of friendships between 15 children in a school and 12 pairs of mutually good friends. The children are categorized by gender, girl (G) or boy (B), and living area, north (N) or south (S).

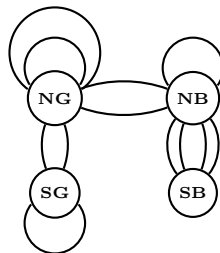


Figure 2: Final multigraph of the friendships in Figure 1 categorized by gender and living area. The edges represent pairwise friendships within and between categories.

As another illustration of vertex aggregation, consider network of co-operations between business companies which are categorized by branch. Figure 3 shows 20 co-operation pairs between 25 companies belonging to branch A, B or C. The multigraph on the three branches is given in Figure 4 and has edge multiplicities that represent co-operations within and between the branches. It is conveniently presented in table format.

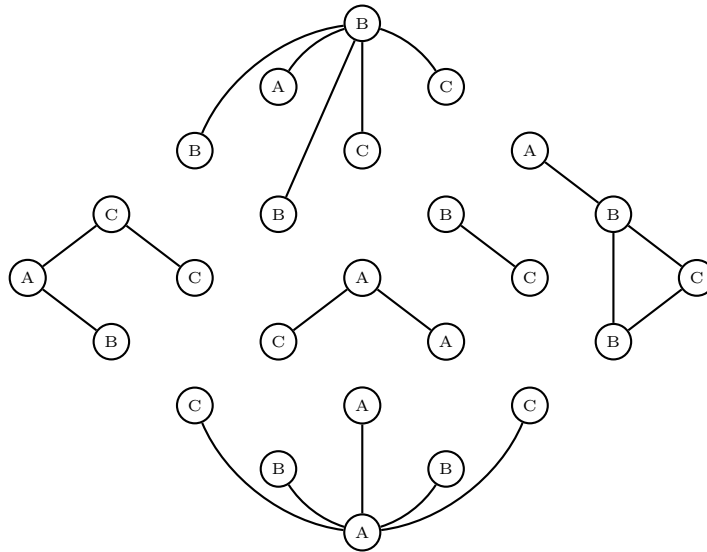


Figure 3: Initial graph of 20 co-operations between 25 companies. The companies are categorized by three different branches labeled A, B and C.

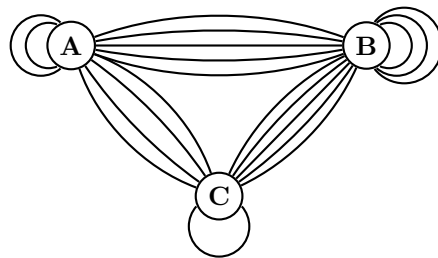


Figure 4: Final multigraph of the co-operations in Figure 3 categorized by branch. The edges represent pairwise co-operations within and between the three branches.

Vertex aggregation is a powerful method to analyze structure in very large graphs. Another type of aggregation that is often useful is edge aggregation, which is illustrated by the following example of a time series. Assume that we are studying a graph with a fixed number of vertices and different categories of pairwise contacts. Further, assume that we study this graph over a period of time, i.e. how the different contacts vary over a time period. For example, let the initial graph have 5 vertices representing 5 different departments in a company and we study the variation of 3 different edge categories representing pairwise contact types between and within the departments. These contact types are phone call, video call or meeting. The connections between the five departments have been observed every day for a total time period of three days. This is illustrated in Figure 5 where the edge attributes are labeled with the colors blue, red and green.

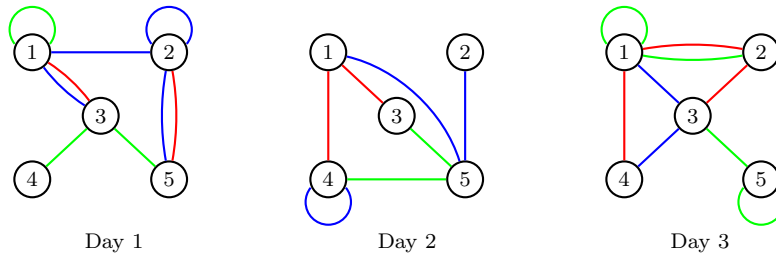


Figure 5: Initial daily graphs on 5 different departments of a company showing three connection types labeled blue (phone call), red (video call) and green (meeting).

The transformation with respect to edge attributes of the graphs in Figure 5 can be done in different ways. If we aggregate over time periods, we obtain for each edge category a multigraph for the total time period of three days, which is shown in Figure 6.

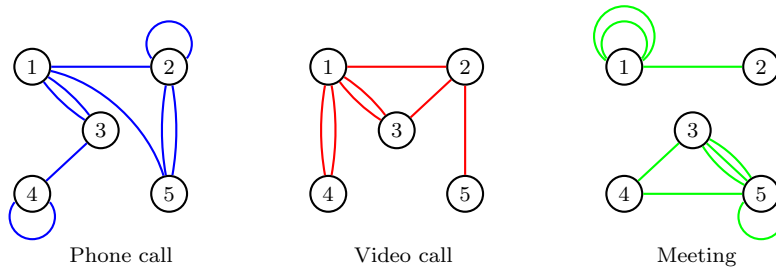


Figure 6: Multigraphs obtained by aggregating each of the edge categories in Figure 5 over all three days.

Another way of transforming the initial graphs in Figure 5 to multigraphs is by aggregating over contact types (ignoring edge colors), to get one multigraph for each time period. If we also aggregate over the three time periods we obtain a multigraph with 5 vertices and a total of 25 edges, shown in Figure 7.

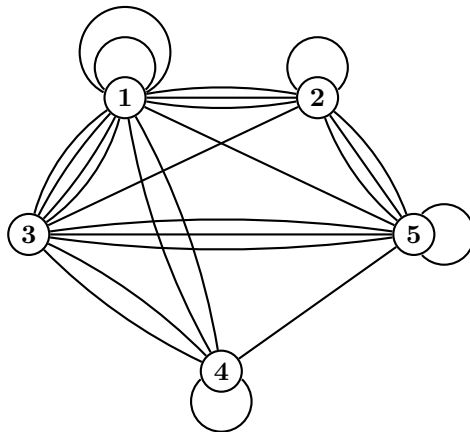


Figure 7: Final multigraph of the total number of connections during 3 days within and between the five departments in Figure 5.

3 Some Random Multigraph Models

In order to analyze multigraphs, we perform statistical tests of some random multigraph models considered in Frank and Shafie (2012) and Shafie (2012). First we introduce some basic notations. A finite graph g with n labeled vertices and m labeled edges associates with each edge an ordered or unordered vertex pair. Let $V = \{1, \dots, n\}$ and $E = \{1, \dots, m\}$ be the sets of vertices and edges labeled by integers, and let R denote the set of available sites for the edges. For directed graphs the site space is $R = V^2$ and the number of sites is given by $r = n^2$. For undirected graphs we use the site space $R = \{(i, j) \in V^2 : i \leq j\}$ where we consider (i, j) with $i \leq j$ as a canonical representation for the unordered vertex pair. The number of sites for undirected graphs is given by $r = \binom{n+1}{2}$. The graph is thus an injective map $g : E \rightarrow R \subseteq V^2$.

A random multigraph is given by a probability distribution over some class of multigraphs. A multigraph with labeled vertices and undistinguished edges is represented by the edge multiplicity sequence $\mathbf{m} = (m_{ij} : (i, j) \in R)$ where the edge multiplicity m_{ij} denotes the number of multiple edges at site $(i, j) \in R$. For undirected multigraphs, the edge sites are listed in the canonical order

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

so that m_{ii} is the number of loops at vertex i , and m_{ij} for $i < j$ is the number of edges between vertices i and j . In this case it is convenient to define $m_{ij} = 0$ for $i > j$. The edge multiplicity sequence \mathbf{m} has total

$$m_{..} = \sum_{i \leq j} m_{ij} = m \quad \text{and} \quad m_{i.} + m_{.i} = \sum_{j=1}^n m_{ij} + \sum_{j=1}^n m_{ji} = d_i$$

is the degree of vertex i , which can also be interpreted as the number of edge-stubs or half-edges at vertex i for $i = 1, \dots, n$. The stub multiplicity sequence $\mathbf{d} = (d_1, \dots, d_n)$ has total $\sum_{i=1}^n d_i = 2m$.

Consider a random undirected multigraph model where the edges are independently assigned to sites according to a common probability model. Let Q_{ij} denote the probability of assigning an edge to site $(i, j) \in R$ so that $\sum_{i \leq j} Q_{ij} = 1$. This independent edge assignment (IEA) model has edge multiplicity sequence $\mathbf{m}(\text{IEA})$ that is multinomially distributed with parameters m and $\mathbf{Q} = (Q_{ij} : (i, j) \in R)$ so that edge sequences \mathbf{m} have probabilities

$$P(\mathbf{m}(\text{IEA}) = \mathbf{m}) = \binom{m}{\mathbf{m}} \mathbf{Q}^{\mathbf{m}} = \frac{m!}{\prod_{i \leq j} m_{ij}!} \prod_{i \leq j} Q_{ij}^{m_{ij}}.$$

Another random multigraph model is obtained by assuming that the edges are formed by random matching of pairs of edge-stubs in a given sequence of edge-stubs. This random stub matching (RSM) model has fixed stub multiplicity sequence $\mathbf{d} = (d_1, \dots, d_n)$. Under RSM, the edge assignments to sites are dependent. The probability that an edge is assigned to site $(i, j) \in R$ is given by

$$Q_{ij} = \begin{cases} \binom{d_i}{2} / \binom{2m}{2} & \text{for } i = j \\ d_i d_j / \binom{2m}{2} & \text{for } i < j, \end{cases}$$

so that the edge probability sequence $\mathbf{Q} = \mathbf{Q}(\mathbf{d})$ is a function of the stub multiplicity sequence \mathbf{d} . The probability of edge multiplicity sequence \mathbf{m} under RSM is shown in Shafie (2012) to be given by

$$P(\mathbf{m}(\text{RSM}) = \mathbf{m}) = \frac{2^{m_2} \binom{m}{\mathbf{m}}}{\binom{2m}{\mathbf{d}}} = \frac{2^{m_2} m! \prod_{i=1}^n d_i!}{(2m)! \prod_{i \leq j} m_{ij}!},$$

where $m_2 = \sum_{i < j} m_{ij}$.

A Bayesian version of the RSM model is obtained by assuming that the stubs are independently assigned to vertices according to a probability distribution $\mathbf{p} = (p_1, \dots, p_n)$. The stub multiplicity sequence under independent stub assignments (ISA) is multinomially distributed with parameters $2m$ and \mathbf{p} . This multinomial distribution can be viewed as a

Bayesian model for the stub multiplicities and leads to independent edge assignments. Thus by the Bayesian assumption the RSM model is turned into a special IEA model with edge probability sequence \mathbf{Q} defined as a function of \mathbf{p} according to

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

Another way to get an approximate IEA model from an RSM model is to ignore the dependency between the edge assignments in the RSM model. The edge probability sequence $\mathbf{Q} = \mathbf{Q}(\mathbf{d})$ of the RSM model is used to define a model with independent edge assignment of stubs (IEAS). Note that the IEAS model, like other IEA models, has $\binom{m+r-1}{m}$ different outcomes of \mathbf{m} , while the RSM models are restricted to outcomes that are consistent with stub multiplicity sequence \mathbf{d} only.

The following notations will be used for the models presented in this section. Independent edge assignment is denoted IEA(\mathbf{Q}), random stub matching is denoted RSM(\mathbf{d}), independent stub assignments is denoted ISA(\mathbf{p}), and independent edge assignments of stubs is denoted IEAS(\mathbf{d}).

4 Statistical Tests of a Simple Multigraph Hypothesis

4.1 Test Statistics

A simple multigraph hypothesis H_0 is defined as a fully specified IEA(\mathbf{Q}_0) which can be an ISA(\mathbf{p}_0) or an IEAS(\mathbf{d}_0) with \mathbf{Q}_0 specified as a function of \mathbf{d}_0 or \mathbf{p}_0 . The tests are performed using goodness-of-fit measures between the multiplicity sequence \mathbf{m} of an observed multigraph and the expected multiplicity sequence according to H_0 .

Asymptotic theory for likelihood ratios and goodness-of-fit statistics is given for instance by Anderson (1980) and Cox and Hinkley (1974). The Pearson goodness-of-fit statistic is given by

$$S_0 = \sum_{i \leq j} \sum_{i \leq j} \frac{(m_{ij} - mQ_{0ij})^2}{mQ_{0ij}} = \sum_{i \leq j} \sum_{i \leq j} \frac{m_{ij}^2}{mQ_{0ij}} - m ,$$

which is asymptotically χ^2 -distributed with $df = r - 1$ degrees of freedom if the multiplicity sequence is obtained according to IEA(\mathbf{Q}) and the correct model $\mathbf{Q}_0 = \mathbf{Q}$ is tested. We denote a random variable with this distribution χ_{r-1}^2 . The divergence statistic is given by

$$D_0 = \sum_{i \leq j} \sum_{i \leq j} \frac{m_{ij}}{m} \log \frac{m_{ij}}{mQ_{0ij}} ,$$

and an asymptotic χ_{r-1}^2 -statistic can be obtained as

$$T_0 = \frac{2m}{\log e} D_0 .$$

Divergence statistics are used as goodness-of-fit statistics for instance by Kullback (1959) and Frank (2011). For good asymptotics it is normally assumed that m is large and mQ_{ij} is not too small (for instance $mQ_{ij} \geq 5$ and $m \geq 5r$). By approximation of the logarithm function it can be shown that $S_0 \approx T_0$ for large m . The critical region for the tests is taken as values of S_0 and T_0 above a critical value cv given by

$$cv = df + 2\sqrt{2df} = r - 1 + \sqrt{8(r-1)} ,$$

which has a significance level approximately equal to 5% given by

$$\alpha = P(\chi_{r-1}^2 > cv) .$$

The power functions

$$P(S_0 > cv) = 1 - \beta_{S_0}(\mathbf{Q}) \quad \text{and} \quad P(T_0 > cv) = 1 - \beta_{T_0}(\mathbf{Q})$$

are calculated using the distributions of S_0 and T_0 when \mathbf{m} is multinomially distributed with parameters m and \mathbf{Q} , for $\mathbf{Q} = \mathbf{Q}_0$ and for $\mathbf{Q} \neq \mathbf{Q}_0$. Specifically, S_0 and T_0 are compared to χ_{r-1}^2 via moments and cumulative distribution functions. For instance, the expected value of S_0 reveals how far from $E(\chi_{r-1}^2) = r - 1$ the distribution of S_0 is. This expected value is given by

$$E(S_0) = \sum_{i \leq j} \sum \frac{E(m_{ij}^2)}{mQ_{0ij}} - m = \sum_{i \leq j} \sum \frac{Q_{ij} + (m-1)Q_{ij}^2}{Q_{0ij}} - m ,$$

where m_{ij} is binomially distributed with parameters m and Q_{ij} so that

$$E(m_{ij}^2) = \text{Var}(m_{ij}) + [E(m_{ij})]^2 = mQ_{ij}(1 - Q_{ij}) + m^2Q_{ij}^2 = mQ_{ij} + m(m-1)Q_{ij}^2 .$$

In particular, if $\mathbf{Q} = \mathbf{Q}_0$ so that $Q_{ij} = Q_{0ij}$ for $i \leq j$, the null distribution of S_0 has expected value

$$E(S_0) = \sum_{i \leq j} \sum [1 + (m-1)Q_{ij}] - m = r - 1 .$$

Under the ISA(\mathbf{p}) model and ISA(\mathbf{p}_0) hypothesis, the expected value of S_0 is given as

$$\begin{aligned} E(S_0) &= \sum_{i=1}^n L_i^2 [1 + (m-1)p_i^2] + \sum_{i \neq j} \sum \frac{L_i L_j}{2} [1 + (m-1)2p_i p_j] - m \\ &= \sum_{i=1}^n L_i^2 + (m-1) \sum_{i=1}^n (L_i p_i)^2 + \sum_{i=1}^n \sum_{j=1}^n \frac{L_i L_j}{2} \\ &\quad + (m-1) \sum_{i=1}^n \sum_{j=1}^n L_i L_j p_i p_j - \sum_{i=1}^n \frac{L_i^2}{2} - (m-1) \sum_{i=1}^n (L_i p_i)^2 - m \\ &= \frac{\sum_{i=1}^n L_i^2 + (\sum_{i=1}^n L_i)^2}{2} - m + (m-1) \left(\sum_{i=1}^n L_i p_i \right)^2 , \end{aligned}$$

where $L_i = p_i/p_{0i}$ is the likelihood ratio for stub assignments. As seen, the variation of $E(S_0)$ depends on $\sum_{i=1}^n L_i$, $\sum_{i=1}^n L_i^2$ and $\sum_{i=1}^n L_i p_i$. In particular, for a uniform ISA(\mathbf{p}_0) hypothesis where $p_{0i} = 1/n$,

$$E(S_0) = \frac{n^2 \sum_{i=1}^n p_i^2 + n^2}{2} - m + (m-1)n^2 \left(\sum_{i=1}^n p_i^2 \right)^2 ,$$

which by letting $s_2 = \sum_{i=1}^n p_i^2$ can be simplified to

$$E(S_0) = m(n^2 s_2^2 - 1) + \frac{n^2}{2}(1 + s_2 - 2s_2^2) .$$

From this we see that $E(S_0)$ grows linearly with m having coefficients depending on n and s_2 . By using

$$E(S_0) = s_2^2 n^2 (m-1) + s_2 \frac{n^2}{2} + \frac{n^2}{2} - m$$

and $1/n \leq s_2 \leq 1$, it follows that

$$r-1 \leq E(S_0) \leq m(n^2 - 1) .$$

We also note that if $\mathbf{p} = \mathbf{p}_0$ so that $p_i = p_{0i}$, the null distribution has

$$E(S_0) = \frac{n + n^2}{2} - m + (m-1) = \binom{n+1}{2} - 1 = r-1$$

which is consistent with the result shown previously for $\mathbf{Q} = \mathbf{Q}_0$.

The expected value of S_0 can also be considered for the RSM(\mathbf{d}) model when H_0 is RSM(\mathbf{d}_0) or IEAS(\mathbf{d}_0) since \mathbf{Q}_0 of IEAS and RSM are identical. Shafie (2012) gives the moments of m_{ij} under RSM as

$$E(m_{ij}) = mQ_{ij} \quad \text{for } i \leq j ,$$

and

$$\text{Var}(m_{ij}) = \sigma_{ij}^2 + \Delta_{ij} \quad \text{for } i \leq j ,$$

where $\sigma_{ij}^2 = mQ_{ij}(1 - Q_{ij})$ is the variance under IEA, and Δ_{ij} is the difference between the variances of m_{ij} under RSM and IEA:

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

where

$$Q_{ijij} = \begin{cases} Q_{ii} \left(\frac{(d_i-2)(d_i-3)}{(2m-2)(2m-3)} \right) & \text{for } i = j \\ Q_{ij} \left(\frac{2(d_i-1)(d_j-1)}{(2m-2)(2m-3)} \right) & \text{for } i < j . \end{cases}$$

A general expression for the expected value of S_0 under RSM is here obtained as

$$\begin{aligned}
E(S_0) &= \sum_{i \leq j} \sum \frac{E(m_{ij}^2)}{mQ_{0ij}} - m \\
&= \sum_{i \leq j} \sum \frac{\sigma_{ij}^2 + \Delta_{ij} + m^2 Q_{ij}^2}{mQ_{0ij}} - m \\
&= \sum_{i \leq j} \sum \frac{mQ_{ij}(1 - Q_{ij}) + \Delta_{ij} + m^2 Q_{ij}^2}{mQ_{0ij}} - m .
\end{aligned}$$

For $\mathbf{Q} = \mathbf{Q}_0$ so that $Q_{ij} = Q_{0ij}$ for $i \leq j$, this simplifies to

$$\begin{aligned}
E(S_0) &= r - 1 + \sum_{i \leq j} \sum \frac{\Delta_{ij}}{mQ_{ij}} \\
&= r - 1 + \sum_{i \leq j} \sum \frac{m(m-1)(Q_{ijij} - Q_{ij}^2)}{mQ_{ij}} \\
&= r - 1 + (m-1) \left[\sum_{i \leq j} \sum \frac{Q_{ijij}}{Q_{ij}} - \sum_{i \leq j} \sum Q_{ij} \right] \\
&= r - m + (m-1) \left[\sum_{i \leq j} \sum \frac{Q_{ijij}}{Q_{ij}} \right] \\
&= r - m + (m-1) \left[\sum_{i < j} \sum \frac{2(d_i - 1)(d_j - 1)}{(2m-2)(2m-3)} + \sum_{i=1}^n \frac{(d_i - 2)(d_i - 3)}{(2m-2)(2m-3)} \right] \\
&= r - m + \frac{1}{2(2m-3)} \left[\sum_{i \neq j} \sum (d_i - 1)(d_j - 1) + \sum_{i=1}^n (d_i - 2)(d_i - 3) \right] \\
&= r - m + \frac{1}{2(2m-3)} [4m^2 + 4mn + n^2 - 6m + 5n] \\
&= \frac{(m-1)n(n-1)}{2m-3} ,
\end{aligned}$$

which implies that the expected value of the null distribution only depends on the number of vertices and edges. Using this expression we can now show for which values of m and n the expected value of S_0 under RSM is smaller than $r - 1$, i.e.

$$E(S_0) = \frac{(m-1)n(n-1)}{2m-3} < (r-1) = \frac{n(n+1)}{2} - 1 .$$

Solving the inequality for m gives the following results:

$$E(S_0) < r - 1 \quad \text{for} \quad m > \frac{n + 6}{4} ,$$

$$E(S_0) = r - 1 \quad \text{if} \quad m = \frac{n + 6}{4} \quad \text{is integer} ,$$

and

$$E(S_0) > r - 1 \quad \text{for} \quad m < \frac{n + 6}{4} .$$

Note that the restriction $2m \geq n$ imposed by no isolated vertices implies that $E(S_0) > r - 1$ only for some degenerate cases ($n = 2, m = 1$) and the extreme cases $n = 3$ or 4 , and $m = 2$. Therefore, under RSM the null distribution of the test statistic S_0 has for all other cases an expected value below $r - 1$, and its cumulative distribution function will tend to lie on or above that of χ^2_{r-1} for all practical useful cases. Exceptional cases with $m < (n + 6)/4$ have so few stubs to be matched that they are unlikely to be useful in practice. Compare the requirement of large m needed for good χ^2 asymptotics. Note however that the test statistics may not have asymptotic χ^2 -distributions under RSM due to dependency between edges.

Any test statistic S , like S_0 or T_0 , can be approximated by an adjusted χ^2 -distribution given by

$$S^* = \frac{\mu}{k} \chi_k^2 ,$$

where $\mu = E(S)$. For any positive integer k the approximation S^* has the same mean as S and a variance given by

$$Var(S^*) = \frac{2\mu^2}{k} .$$

Two particular approximations S' and S'' are given by S^* for k chosen as the integer part of μ and for $k = r - 1$, respectively. Their variances are

$$Var(S') = \frac{2\mu^2}{\lfloor \mu \rfloor} \quad \text{and} \quad Var(S'') = \frac{2\mu^2}{r - 1} ,$$

and the preferred approximation is the one with variance closest to $Var(S) = \sigma^2$. Equivalently, the preferred adjusted χ^2 -distribution is the one with degrees of freedom closest to $2\mu^2/\sigma^2$. A good approximation is useful for power calculations.

4.2 Test Illustrations for IEAS Models

We consider multigraphs with 4 vertices and 10 edges and test IEAS(\mathbf{d}_0) hypotheses against IEAS(\mathbf{d}) models. The degree sequences are chosen to include both skew and flat (uniform

and close to uniform) cases. The number of edge sites is here given by $r = 10$ and the test statistics S_0 and T_0 are thus asymptotically χ_9^2 -distributed when the correct model with $\mathbf{d}_0 = \mathbf{d}$ is being tested. The critical value is $cv = 17.49$ and $\alpha = P(\chi_9^2 > cv) = 0.04$. The powers of these tests according to S_0 and T_0 are given in Table 1, where the diagonal representing $\mathbf{d}_0 = \mathbf{d}$ is shaded. Note that there is one case where the order between the components in \mathbf{d}_0 is switched. For this special case, the large deviations between the degree values in models and hypotheses result in powers being close or equal to one for both statistics. When $\mathbf{d}_0 = \mathbf{d}$, $\alpha_{T_0} = 1 - \beta_{T_0} < \alpha \leq 1 - \beta_{S_0} = \alpha_{S_0}$. For flat $\mathbf{d}_0 = \mathbf{d}$, both statistics have significance levels equal or close to α , but for skew $\mathbf{d}_0 = \mathbf{d}$, the significance level of T_0 is much below α and that of S_0 is much above α . For the majority of cases with not too skew $\mathbf{d}_0 \neq \mathbf{d}$, both statistics have fairly good powers, but the inequalities between them persist indicating that their cumulative distribution functions can approach an asymptotic distribution from either below or above. To illustrate the fit of the distributions of the statistics S_0 and T_0 to χ_9^2 , their cumulative distribution functions are shown in Figure 8. For flat $\mathbf{d}_0 = \mathbf{d}$, the null distribution of S_0 almost coincides with that of χ_9^2 . For skew $\mathbf{d}_0 = \mathbf{d}$, the null distributions of both statistics give poor fit to χ_9^2 -distribution. This poor fit is also noted for both flat and skew $\mathbf{d}_0 \neq \mathbf{d}$. Both S_0 and T_0 seem to have distributions that would be better approximated by χ^2 with degrees of freedom chosen to be higher than $r - 1$ in cases with $\mathbf{d}_0 \neq \mathbf{d}$.

The speed of the convergence of the cumulative distribution functions of S_0 and T_0 is illustrated in Figures 9 and 10 where both flat and skew $\mathbf{d}_0 = \mathbf{d}$ are considered. The number of edges m increases as multiples of the chosen degree sequences. We see that even for small m , the null distributions of both statistics are fairly well approximated by their asymptotic χ^2 -distribution. A similar investigation of the non-null distributions of S_0 and T_0 is shown in Figure 11 for flat $\mathbf{d}_0 \neq \mathbf{d}$ and in Figure 12 for skew $\mathbf{d}_0 \neq \mathbf{d}$, where \mathbf{d}_0 is kept fixed and \mathbf{d} is varied. For both flat and skew \mathbf{d}_0 , the deviations between the non-null distributions of S_0 and T_0 and their asymptotic null distribution increase with the number of edges, and even for $m = 12$ this deviation is clearly notable. Thus even for the rather small $m = 12$, it is easy to detect simple hypotheses about false models.

Two cases in Table 2 illustrate how test statistics can be approximated by adjusted χ^2 -distributions. The approximated goodness-of-fit statistics are S'_0 and S''_0 , and the approximated divergence statistics are T'_0 and T''_0 . These approximations are evaluated by comparing their variances to $Var(S_0)$ and $Var(T_0)$. The expected values and variances of all versions of the test statistics are presented in Table 2 where the versions that are not preferred are shaded so that it is easier to compare preferences in different cases. For the first case, S''_0 is preferred to S'_0 while T'_0 is preferred to T''_0 . Equivalently, the preferred adjusted χ^2 -distribution for S_0 has $df = r - 1 = 9$ since it is closer than $df = \lfloor \mu \rfloor = 13$ to $2E(S_0)/Var(S_0) = 7.31$, and the adjusted χ^2 -distribution for T_0 has $df = \lfloor \mu \rfloor = 13$ since it is closer than $df = r - 1 = 9$ to $2E(T_0)/Var(T_0) = 18.19$. For the second case, S'_0 is preferred to S''_0 , while T''_0 is preferred to T'_0 .

Table 1: Power according to S_0 (upper value) and T_0 (value below) when model is IEAS(\mathbf{d}) and hypothesis is IEAS(\mathbf{d}_0) for $n = 4$ and $m = 10$. $\alpha = 0.04$.

\mathbf{d}_0	\mathbf{d}						
	(14, 2, 2, 2)	(12, 3, 3, 2)	(9, 7, 2, 2)	(8, 8, 2, 2)	(6, 6, 6, 2)	(6, 5, 5, 4)	(5, 5, 5, 5)
(14, 2, 2, 2)	0.19 0.01	0.42 0.06	0.87 0.52	0.94 0.70	0.98 0.87	0.97 0.82	0.99 0.92
(2, 2, 14, 2)	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	0.98 0.87	0.99 0.92	0.99 0.92
(12, 3, 3, 2)	0.06 0.01	0.10 0.01	0.50 0.24	0.66 0.40	0.75 0.52	0.77 0.50	0.89 0.69
(9, 7, 2, 2)	0.34 0.25	0.31 0.14	0.13 0.01	0.14 0.02	0.74 0.40	0.78 0.44	0.87 0.60
(8, 8, 2, 2)	0.58 0.47	0.44 0.26	0.13 0.02	0.13 0.01	0.73 0.38	0.78 0.44	0.87 0.58
(6, 6, 6, 2)	0.85 0.74	0.54 0.39	0.24 0.19	0.21 0.18	0.08 0.02	0.36 0.14	0.53 0.27
(6, 5, 5, 4)	0.78 0.69	0.46 0.36	0.26 0.22	0.28 0.23	0.07 0.06	0.05 0.03	0.07 0.05
(5, 5, 5, 5)	0.91 0.63	0.70 0.63	0.48 0.43	0.46 0.41	0.13 0.12	0.05 0.04	0.04 0.03

Table 2: Moments of S_0 , S'_0 , S''_0 , T_0 , T'_0 and T''_0 for some IEAS(\mathbf{d}) models and IEAS(\mathbf{d}_0) hypotheses with $n = 4$ and $m = 10$. The unshaded columns correspond to the best approximations to the test statistics.

Case 1: $\mathbf{d}_0 = (6, 6, 6, 2)$, $\mathbf{d} = (8, 8, 2, 2)$						
	S_0	S'_0	S''_0	T_0	T'_0	T''_0
Expected value	13.62	13.62	13.62	13.40	13.40	13.40
Variance	50.74	28.52	41.20	19.74	27.64	39.93
Case 2: $\mathbf{d}_0 = (12, 3, 3, 2)$, $\mathbf{d} = (14, 2, 2, 2)$						
	S_0	S'_0	S''_0	T_0	T'_0	T''_0
Expected value	7.51	7.51	7.51	7.82	7.82	7.82
Variance	31.79	16.14	12.55	10.61	17.47	13.59

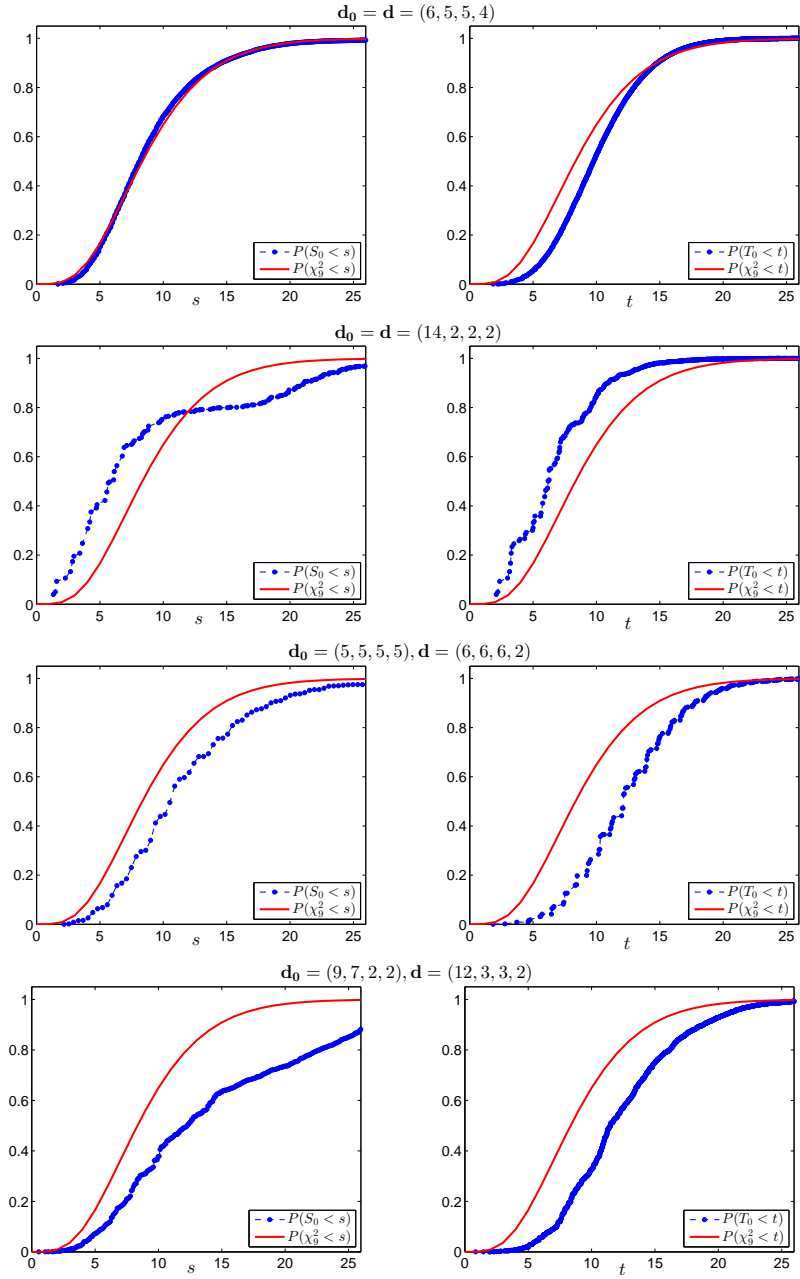


Figure 8: Distributions of S_0 , T_0 , and χ_9^2 for some IEAS(\mathbf{d}) models and IEAS(\mathbf{d}_0) hypotheses with $n = 4$ and $m = 10$.

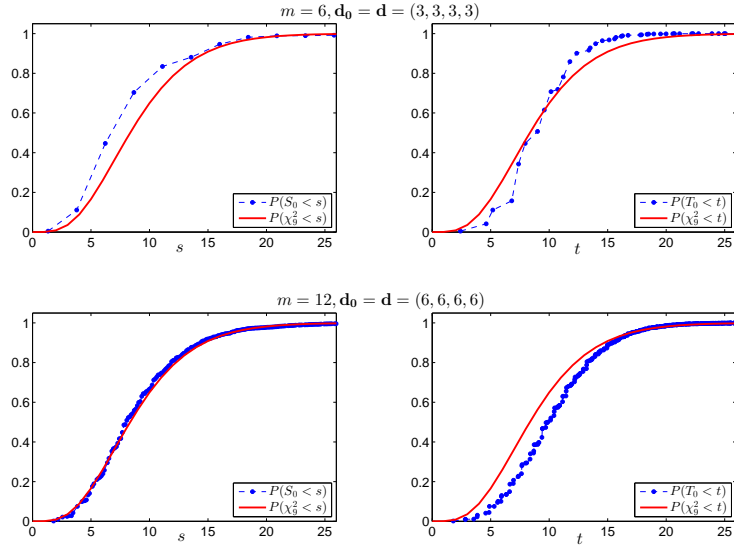


Figure 9: Null distributions of S_0 and T_0 for some IEAS(\mathbf{d}) models and IEAS(\mathbf{d}_0) hypotheses with flat $\mathbf{d}_0 = \mathbf{d}$ when m increases.

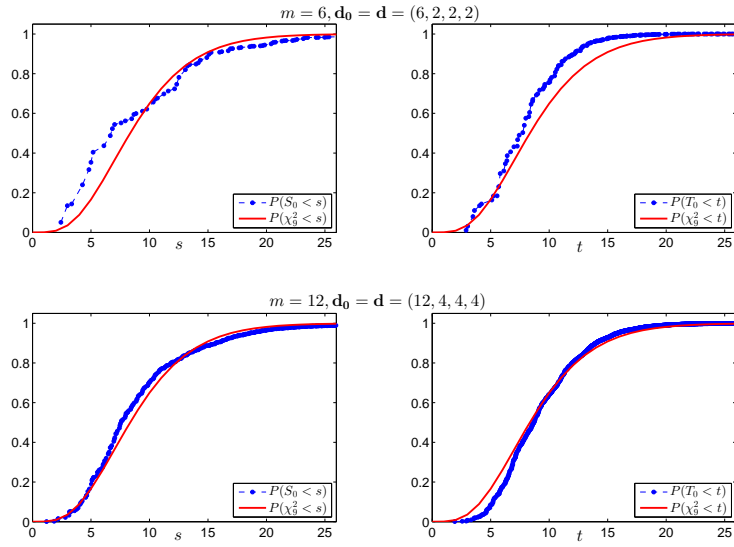


Figure 10: Null distributions of S_0 and T_0 for some IEAS(\mathbf{d}) models and IEAS(\mathbf{d}_0) hypotheses with skew $\mathbf{d}_0 = \mathbf{d}$ when m increases.

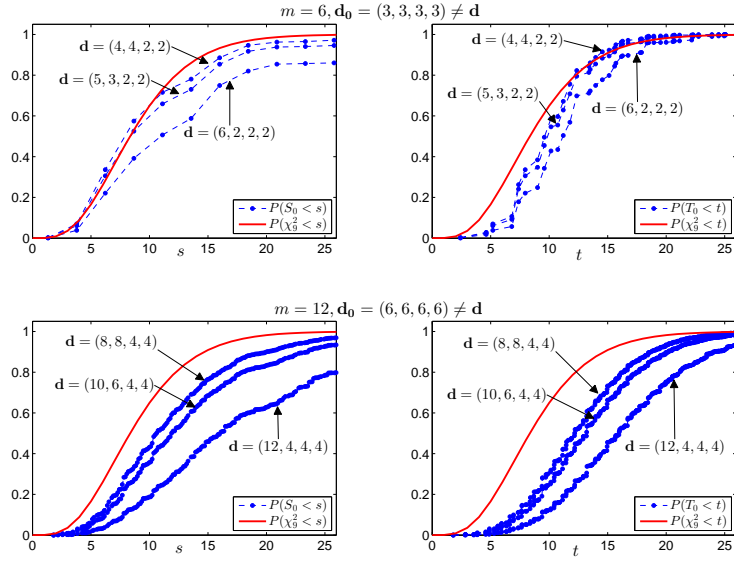


Figure 11: Non-null distributions of S_0 and T_0 for some IEAS(\mathbf{d}) models and IEAS(\mathbf{d}_0) hypotheses with flat \mathbf{d}_0 and different \mathbf{d} when m increases.

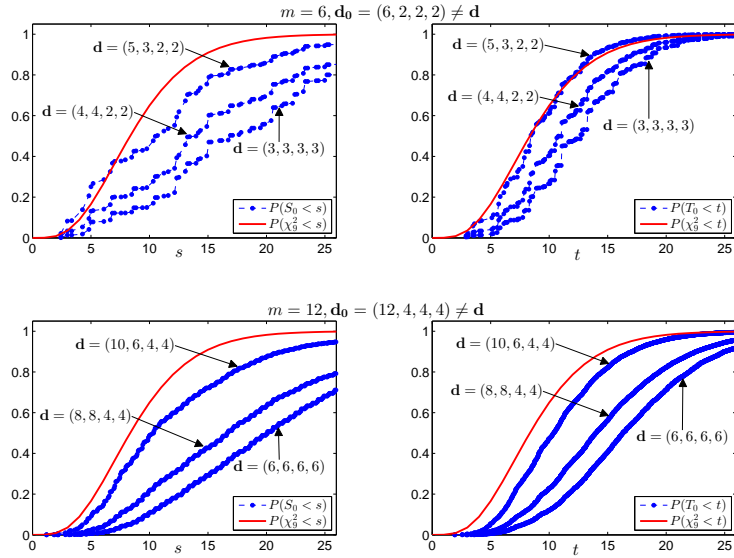


Figure 12: Non-null distributions of S_0 and T_0 for some IEAS(\mathbf{d}) models and IEAS(\mathbf{d}_0) hypotheses with skew \mathbf{d}_0 and different \mathbf{d} when m increases.

4.3 Test Illustrations for ISA Models

We now turn our attention to $\text{ISA}(\mathbf{p}_0)$ hypotheses tested against $\text{ISA}(\mathbf{p})$ models and consider tests of different multigraphs of the same size. The stub selection probability sequences are chosen to illustrate both skew and flat cases. Note that there is one case where the order between the components in \mathbf{p}_0 is switched. The powers of these tests according to S_0 and T_0 are given in Table 3. We see that most results are consistent with those seen in Table 1 for IEAS models. For $\mathbf{p}_0 = \mathbf{p}$, α_{S_0} and α_{T_0} are on opposite sides of $\alpha = 0.04$ but they are both close to α except for very skew cases. For the majority of cases with $\mathbf{p}_0 \neq \mathbf{p}$, both test statistics have reasonable powers unless \mathbf{p}_0 and \mathbf{p} are too close. In Figure 13, the fit of the distributions of the statistics S_0 and T_0 to that of χ_9^2 are illustrated for some selected cases. Overall, we see that even for these examples with small m , we have fairly good fit for all illustrated cases with both flat and skew $\mathbf{p}_0 = \mathbf{p}$, and $\mathbf{p}_0 \neq \mathbf{p}$.

The impact on the null and non-null distributions of S_0 and T_0 for skew and flat \mathbf{p}_0 when m increases is illustrated in Figures 14 to 17 where similar results as those for IEAS models are noted. The convergence to the asymptotic distribution is rapid for null distributions of both statistics, and the deviations between the non-null distributions of both statistics and their asymptotic null distribution increase with m . The latter result implies that adjusted χ^2 -distributions should be used to approximate the non-null distributions.

Two cases from Table 3 are chosen to illustrate the performance of the approximate test statistics. By comparing variances we obtain the results presented in Table 4, where non-preferred statistics are shaded.

Table 3: Power according to S_0 (upper value) and T_0 (value below) when model is $\text{ISA}(\mathbf{p})$ and hypothesis is $\text{ISA}(\mathbf{p}_0)$ for $n = 4$ and $m = 10$. $\alpha = 0.04$.

\mathbf{p}_0	\mathbf{p}					
	$(\frac{7}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$	$(\frac{3}{5}, \frac{1}{5}, \frac{1}{10}, \frac{1}{10})$	$(\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$	$(\frac{4}{9}, \frac{1}{3}, \frac{1}{9}, \frac{1}{9})$	$(\frac{3}{8}, \frac{3}{8}, \frac{1}{8}, \frac{1}{8})$	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
$(\frac{7}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$	0.10 0.01	0.33 0.07	0.57 0.19	0.80 0.47	0.92 0.67	0.99 0.88
$(\frac{1}{10}, \frac{1}{10}, \frac{7}{10}, \frac{1}{10})$	1.00 1.00	1.00 1.00	1.00 0.99	1.00 1.00	1.00 1.00	0.99 0.88
$(\frac{3}{5}, \frac{1}{5}, \frac{1}{10}, \frac{1}{10})$	0.06 0.01	0.08 0.01	0.36 0.11	0.32 0.11	0.50 0.24	0.92 0.72
$(\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$	0.04 0.05	0.04 0.03	0.06 0.02	0.27 0.14	0.42 0.25	0.53 0.35
$(\frac{4}{9}, \frac{1}{3}, \frac{1}{9}, \frac{1}{9})$	0.27 0.24	0.09 0.05	0.29 0.14	0.06 0.02	0.11 0.04	0.74 0.49
$(\frac{3}{8}, \frac{3}{8}, \frac{1}{8}, \frac{1}{8})$	0.54 0.47	0.19 0.14	0.29 0.19	0.04 0.02	0.05 0.02	0.58 0.38
$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	0.88 0.86	0.66 0.63	0.32 0.28	0.35 0.33	0.25 0.23	0.03 0.03

Table 4: Moments of S_0 , S'_0 , S''_0 , T_0 , T'_0 and T''_0 for some $\text{ISA}(\mathbf{p})$ models and $\text{ISA}(\mathbf{p}_0)$ hypotheses with $n = 4$ and $m = 10$. The unshaded columns correspond to the best approximations to the test statistics.

Case 1: $\mathbf{p}_0 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $\mathbf{p} = (\frac{3}{8}, \frac{3}{8}, \frac{1}{8}, \frac{1}{8})$						
	S_0	S'_0	S''_0	T_0	T'_0	T''_0
Expected value	14.56	14.56	14.56	14.43	14.43	14.43
Variance	50.98	30.30	47.13	22.05	29.74	46.27
Case 2: $\mathbf{p}_0 = (\frac{3}{5}, \frac{1}{5}, \frac{1}{10}, \frac{1}{10})$, $\mathbf{p} = (\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$						
	S_0	S'_0	S''_0	T_0	T'_0	T''_0
Expected value	17.08	17.08	17.08	11.20	11.20	11.20
Variance	167.82	34.33	64.85	25.83	22.82	27.89

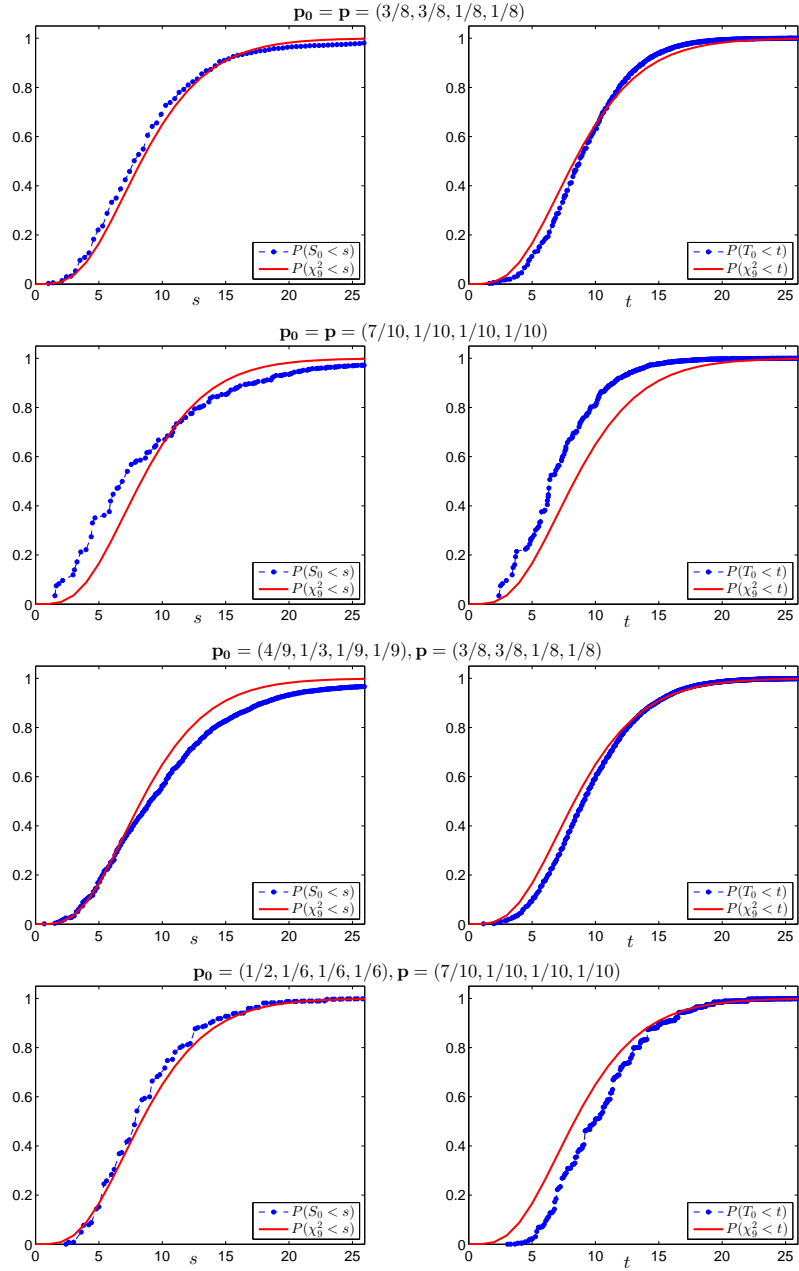


Figure 13: Distributions of S_0 , T_0 and χ_9^2 for some $\text{ISA}(\mathbf{p})$ models and $\text{ISA}(\mathbf{p}_0)$ hypotheses with $n = 4$ and $m = 10$.

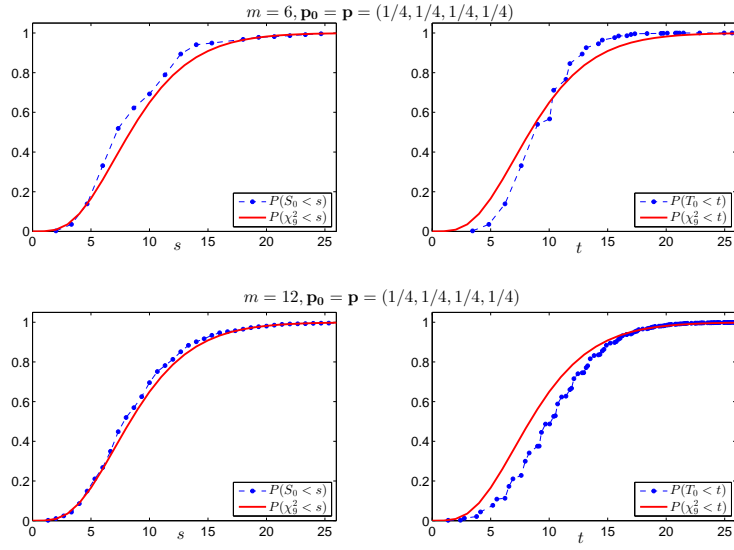


Figure 14: Null distributions of S_0 and T_0 for some $\text{ISA}(\mathbf{p})$ models and $\text{ISA}(\mathbf{p}_0)$ hypotheses with flat $\mathbf{p}_0 = \mathbf{p}$ when m increases.

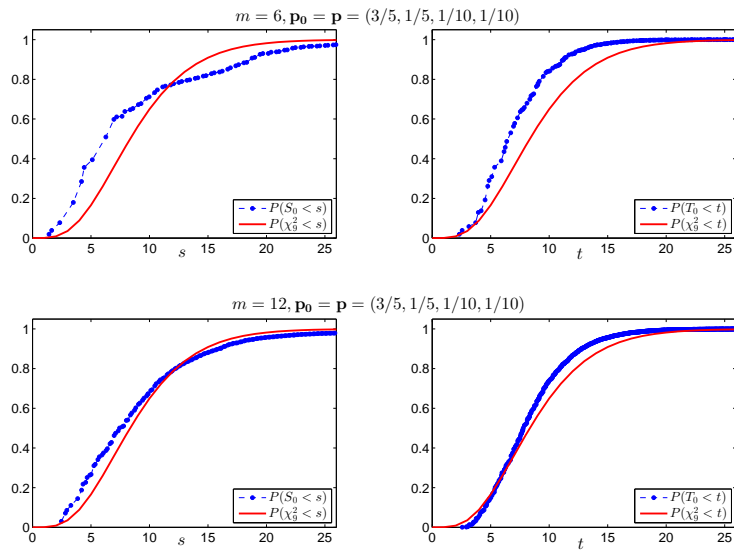


Figure 15: Null distributions of S_0 and T_0 for some $\text{ISA}(\mathbf{p})$ models and $\text{ISA}(\mathbf{p}_0)$ hypotheses with skew $\mathbf{p}_0 = \mathbf{p}$ when m increases.

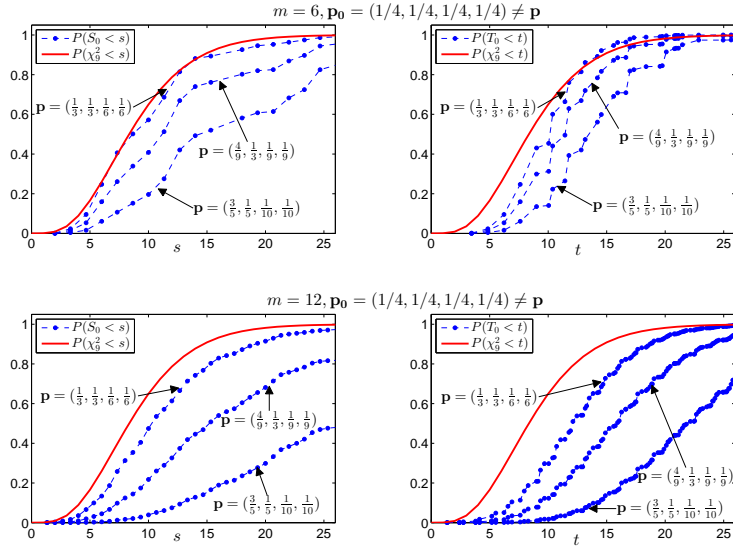


Figure 16: Non-null distributions of S_0 and T_0 for some ISA(\mathbf{p}) models and ISA(\mathbf{p}_0) hypotheses with flat \mathbf{p}_0 and different \mathbf{p} when m increases.

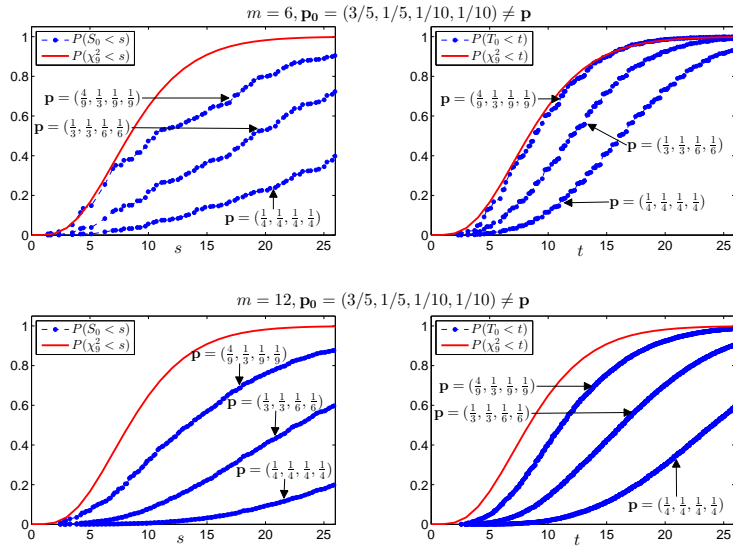


Figure 17: Non-null distributions of S_0 and T_0 for some ISA(\mathbf{p}) models and ISA(\mathbf{p}_0) hypotheses with skew \mathbf{p}_0 and different \mathbf{p} when m increases.

4.4 Test Illustrations for RSM Models

When performing tests of IEA models, multigraphs are known to have multiplicity sequences that are multinomially distributed, which implies that the distributions of the test statistics S_0 and T_0 are asymptotically χ^2 -distributed when the correct model is being tested. However, for RSM models, there is dependence between edges, and the distributions of S_0 and T_0 are unknown. In this section, we illustrate some of the consequences of using the previously described tests of simple hypotheses against a false IEA model when the true model is RSM. Here, both IEAS(\mathbf{d}_0) and ISA(\mathbf{p}_0) hypotheses are tested for flat and skew \mathbf{d}_0 and \mathbf{p}_0 . The true model is RSM(\mathbf{d}) so that only non-null distributions of S_0 and T_0 are considered.

We start by testing multigraphs with 4 vertices and 12 edges. The powers of these tests according to S_0 and T_0 are presented in Table 5. For IEAS(\mathbf{d}_0) hypotheses, the diagonal representing $\mathbf{d}_0 = \mathbf{d}$ is shaded, and for ISA(\mathbf{p}_0) hypotheses, the diagonal representing $\mathbf{p}_0 = \mathbf{d}/2m$ is shaded. For these shaded cases, both α_{S_0} and α_{T_0} are much below $\alpha = 0.04$, except for the very skew $\mathbf{d}_0 = \mathbf{d} = (18, 2, 2, 2)$ where α_{S_0} is much above α . For the majority of cases with $\mathbf{d}_0 \neq \mathbf{d}$ or $\mathbf{p}_0 \neq \mathbf{d}/2m$ both test statistics have good or reasonable powers, unless \mathbf{d} is too close to \mathbf{d}_0 or $2m\mathbf{p}_0$. To illustrate the fit of the distributions of the statistics S_0 and T_0 to that of χ_9^2 , their cumulative distribution functions for some selected cases are shown in Figure 18. We see similar trends as those for IEAS models in Figure 8 and ISA models in Figure 13 which generally makes it hard to detect differences between how the models RSM, IEAS and ISA effect the test statistics.

Four cases are chosen to illustrate adjusted χ^2 -approximations to the distributions of the test statistics. Table 6 shows the expected values and variances for test statistics and approximations, and the approximations that are not preferred are shaded. Thus we see that the preferences can vary in all different ways.

Let us now increase the number of edges and consider multigraphs with 3 vertices and 45 edges. The powers of these tests according to S_0 and T_0 are presented in Table 7 where the following is noted. For IEAS and ISA hypotheses with $\mathbf{d}_0 = \mathbf{d}$ and $\mathbf{p}_0 = \mathbf{d}/2m$, the significance levels of both S_0 and T_0 are much smaller than α and also equal or almost equal. There are some cases of powers below α implying that it is difficult to detect hypotheses about wrong models. For IEAS hypotheses where $\mathbf{d}_0 \neq \mathbf{d}$, and ISA hypotheses where $\mathbf{p}_0 \neq \mathbf{d}/2m$, the powers are equal or close to one another in the majority of cases. This is a consequence of similarities between IEAS and ISA models for large m . Other results concerning the powers in Table 7 are similar to those seen in Table 5. Figure 19 illustrates the fit of the distributions of S_0 and T_0 to that of χ_5^2 for some different cases with $m = 45$ and should be compared to Figure 18 which illustrates $m = 12$. We see strong deviations from χ_5^2 for both S_0 and T_0 , and for flat \mathbf{d}_0 , the distributions are either to the left of the χ_5^2 -distribution or close to it. This is further illustrated in Figures 20-23 and is in contrast to the finding for IEAS models in Figure 9-12 and for ISA models in Figures 14-17.

In Figures 20 to 23, the non-null distribution of S_0 and T_0 for some $\text{RSM}(\mathbf{d})$ models are illustrated where m increases as multiples of different specified \mathbf{d} . Figures 20 and 21 illustrate IEAS hypotheses with flat and skew \mathbf{d}_0 , and Figures 22 and 23 illustrate ISA hypotheses with flat and skew \mathbf{p}_0 . When $\mathbf{d}_0 = \mathbf{d}$ or $\mathbf{p}_0 = \mathbf{d}/2m$, the non-null distributions of both S_0 and T_0 lie above the asymptotic null distributions. This is consistent with results shown in Section 5.1. Further for these cases we see that as m increases, these distributions still lie above the asymptotic null distribution, and a χ^2 -distribution with lower degrees of freedom seem to better approximate these distributions. For cases with $\mathbf{d}_0 \neq \mathbf{d}$ or $\mathbf{p}_0 \neq \mathbf{d}/2m$, the non-null distributions of both statistics move further away from the asymptotic null distribution as m increases, implying a need to use adjusted χ^2 -distributions for better fit.

Three cases to illustrate the approximations by adjusted χ^2 -distributions are given in Table 8. For all three cases, S'_0 and T'_0 are preferred. For the second and third case with $\mathbf{d}_0 = \mathbf{d}$, the variances of S_0 and T_0 are roughly twice their expected values which are approximately equal to 3. Thus, the adjusted χ^2 -distribution for both test statistics seem to be closer to $r - n$ rather than $r - 1$ degrees of freedom under RSM. This is also supported by the expected value of S_0 which according to the result in Section 4.1 is $(m - 1)n(n - 1)/(2m - 3)$ which is about $r - n = n(n - 1)/2$.

So far in this section we have considered the consequences of replacing IEA models with RSM models, but have only tested IEA hypotheses. We conclude this section with a comment about testing RSM hypotheses. A simple $\text{RSM}(\mathbf{d}_0)$ hypothesis has the same \mathbf{Q}_0 as the $\text{IEAS}(\mathbf{d}_0)$ hypothesis, and S_0 and T_0 can not distinguish between these two hypotheses. Should the model be $\text{RSM}(\mathbf{d})$, there is a dependency between edges when they are assigned to sites, which could be used to distinguish between the two hypotheses. This requires a test not using S_0 or T_0 , but a test using the full potential of \mathbf{m} having as its critical region the set $\overline{M(\mathbf{d}_0)}$ consisting of all outcomes \mathbf{m} that are not compatible with \mathbf{d}_0 . This test has zero probability of false rejection of $\text{RSM}(\mathbf{d}_0)$, and its power can be determined as the sum of the probabilities according to $\text{RSM}(\mathbf{d})$ of the outcomes in the critical region. Shafie (2012) gives the $\text{RSM}(\mathbf{d})$ probabilities and specifies outcomes of \mathbf{m} compatible with a fixed degree sequence. We will not pursue details of this test further here.

Table 5: Power according to S_0 (upper value) and T_0 (value below) when model is $\text{RSM}(\mathbf{d})$ and hypothesis is $\text{IEAS}(\mathbf{d}_0)$ or $\text{ISA}(\mathbf{p}_0)$ for $n = 4$ and $m = 12$. $\alpha = 0.04$.

\mathbf{d}_0	\mathbf{d}						
	(18, 2, 2, 2)	(16, 3, 3, 2)	(13, 5, 4, 2)	(8, 8, 4, 4)	(7, 7, 7, 3)	(7, 6, 6, 5)	(6, 6, 6, 6)
(18, 2, 2, 2)	0.14 0.00	0.33 0.00	0.74 0.08	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00
(16, 3, 3, 2)	0.05 0.00	0.05 0.00	0.15 0.01	1.00 0.79	1.00 0.96	1.00 0.95	1.00 1.00
(13, 5, 4, 2)	0.05 0.02	0.04 0.00	0.05 0.00	0.47 0.10	0.49 0.14	0.79 0.28	0.99 0.70
(8, 8, 4, 4)	1.00 1.00	0.57 0.36	0.07 0.03	0.01 0.01	0.07 0.03	0.09 0.03	0.15 0.04
(7, 7, 7, 3)	1.00 1.00	1.00 1.00	0.15 0.07	0.02 0.02	0.00 0.01	0.05 0.02	0.12 0.04
(7, 6, 6, 5)	1.00 1.00	1.00 1.00	0.15 0.14	0.01 0.02	0.01 0.01	0.00 0.01	0.01 0.01
(6, 6, 6, 6)	1.00 1.00	1.00 1.00	0.52 0.37	0.02 0.02	0.02 0.02	0.01 0.01	0.01 0.01
\mathbf{p}_0							
$(\frac{3}{4}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12})$	0.02 0.00	0.00 0.00	0.79 0.04	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00
$(\frac{2}{3}, \frac{1}{8}, \frac{1}{8}, \frac{1}{12})$	0.01 0.00	0.01 0.00	0.15 0.00	0.98 0.77	1.00 0.92	1.00 0.95	1.00 0.99
$(\frac{13}{24}, \frac{5}{24}, \frac{1}{6}, \frac{1}{12})$	0.02 0.02	0.02 0.02	0.01 0.00	0.42 0.09	0.43 0.07	0.78 0.27	0.99 0.76
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6})$	1.00 1.00	1.00 1.00	0.02 0.03	0.00 0.00	0.03 0.02	0.04 0.03	0.08 0.04
$(\frac{7}{24}, \frac{7}{24}, \frac{7}{24}, \frac{1}{8})$	1.00 1.00	0.83 1.00	0.11 0.06	0.02 0.02	0.00 0.01	0.05 0.02	0.12 0.04
$(\frac{7}{24}, \frac{1}{4}, \frac{1}{4}, \frac{5}{24})$	1.00 1.00	0.83 0.84	0.11 0.08	0.01 0.02	0.01 0.01	0.00 0.00	0.01 0.01
$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	1.00 1.00	1.00 1.00	0.52 0.34	0.02 0.02	0.01 0.02	0.00 0.01	0.00 0.01

Table 6: Moments of $S_0, S'_0, S''_0, T_0, T'_0$ and T''_0 for some RSM(\mathbf{d}) models and IEAS(\mathbf{d}_0) or ISA($\mathbf{d}_0/2m$) hypotheses with $n = 4$ and $m = 12$. The unshaded columns correspond to the best approximations to the test statistics.

Case 1: ISA $\mathbf{d}_0 = \mathbf{d} = (7, 6, 6, 5)$						
	S_0	S'_0	S''_0	T_0	T'_0	T''_0
Expected value	6.18	6.18	6.18	7.90	7.90	7.90
Variance	8.70	12.71	8.47	11.32	17.83	13.87
Case 2: IEAS $\mathbf{d}_0 = \mathbf{d} = (16, 3, 3, 2)$						
	S_0	S'_0	S''_0	T_0	T'_0	T''_0
Expected value	6.29	6.29	6.29	5.08	5.08	5.08
Variance	32.41	13.17	8.78	7.25	10.32	5.73
Case 3: ISA $\mathbf{d}_0 = (13, 5, 4, 2), \mathbf{d} = (8, 8, 4, 4)$						
	S_0	S'_0	S''_0	T_0	T'_0	T''_0
Expected value	16.98	16.98	16.98	12.71	12.71	12.71
Variance	39.45	36.02	64.04	10.84	26.93	35.91
Case 4: IEAS $\mathbf{d}_0 = (7, 7, 7, 3), \mathbf{d} = (6, 6, 6, 6)$						
	S_0	S'_0	S''_0	T_0	T'_0	T''_0
Expected value	13.28	13.28	13.28	10.85	10.85	10.85
Variance	47.97	27.12	39.17	11.73	23.53	26.14

Table 7: Power according to S_0 (upper value) and T_0 (value below) when model is RSM(\mathbf{d}) and hypothesis is IEAS(\mathbf{d}_0) or ISA(\mathbf{p}_0) for $n = 3$ and $m = 45$. $\alpha = 0.05$.

\mathbf{d}_0	\mathbf{d}						
	(70, 10, 10)	(65, 15, 10)	(50, 20, 20)	(45, 35, 10)	(40, 30, 20)	(35, 30, 25)	(30, 30, 30)
(70, 10, 10)	0.01 0.01	0.13 0.03	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00
(65, 15, 10)	0.01 0.01	0.01 0.01	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00
(50, 20, 20)	1.00 1.00	0.85 1.00	0.01 0.02	1.00 1.00	0.37 0.23	0.97 0.90	1.00 1.00
(45, 35, 10)	1.00 1.00	1.00 1.00	1.00 1.00	0.01 0.01	1.00 0.56	1.00 1.00	1.00 1.00
(40, 30, 20)	1.00 1.00	1.00 1.00	0.12 0.24	0.13 0.32	0.01 0.01	0.04 0.03	0.42 0.28
(35, 30, 25)	1.00 1.00	1.00 1.00	0.92 0.90	1.00 1.00	0.02 0.03	0.01 0.01	0.03 0.03
(30, 30, 30)	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	0.23 0.25	0.02 0.03	0.01 0.01
\mathbf{p}_0							
$(\frac{7}{9}, \frac{1}{9}, \frac{1}{9})$	0.01 0.01	0.12 0.03	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00
$(\frac{13}{18}, \frac{1}{6}, \frac{1}{9})$	0.01 0.01	0.01 0.01	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00
$(\frac{5}{9}, \frac{2}{9}, \frac{2}{9})$	1.00 1.00	0.85 1.00	0.01 0.02	1.00 1.00	0.34 0.21	0.91 0.84	1.00 1.00
$(\frac{1}{2}, \frac{7}{18}, \frac{1}{9})$	1.00 1.00	1.00 1.00	1.00 1.00	0.01 0.01	1.00 0.54	1.00 1.00	1.00 1.00
$(\frac{4}{9}, \frac{1}{3}, \frac{2}{9})$	1.00 1.00	1.00 1.00	0.11 0.24	0.13 0.32	0.01 0.01	0.04 0.03	0.39 0.28
$(\frac{7}{18}, \frac{1}{3}, \frac{5}{18})$	1.00 1.00	1.00 1.00	0.90 0.90	1.00 1.00	0.02 0.03	0.01 0.01	0.03 0.02
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	0.19 0.24	0.02 0.03	0.01 0.01

Table 8: Moments of $S_0, S'_0, S''_0, T_0, T'_0$ and T''_0 for some RSM(\mathbf{d}) models and IEAS(\mathbf{d}_0) or ISA($\mathbf{d}_0/2m$) hypotheses with $n = 3$ and $m = 45$. The unshaded columns correspond to the best approximations to the test statistics.

Case 1: IEAS $\mathbf{d}_0 = (70, 10, 10), \mathbf{d} = (65, 15, 10)$						
	S_0	S'_0	S''_0	T_0	T'_0	T''_0
Expected value	7.43	7.43	7.43	6.05	6.05	6.05
Variance	15.96	15.77	22.07	5.28	12.19	14.63

Case 2: ISA $\mathbf{d}_0 = \mathbf{d} = (50, 20, 20)$						
	S_0	S'_0	S''_0	T_0	T'_0	T''_0
Expected value	3.01	3.01	3.01	3.32	3.32	3.32
Variance	5.50	6.05	3.63	7.45	7.35	4.41

Case 3: IEAS $\mathbf{d}_0 = \mathbf{d} = (30, 30, 30)$						
	S_0	S'_0	S''_0	T_0	T'_0	T''_0
Expected value	3.03	3.03	3.03	3.14	3.14	3.14
Variance	5.83	6.14	3.68	6.66	6.57	3.94

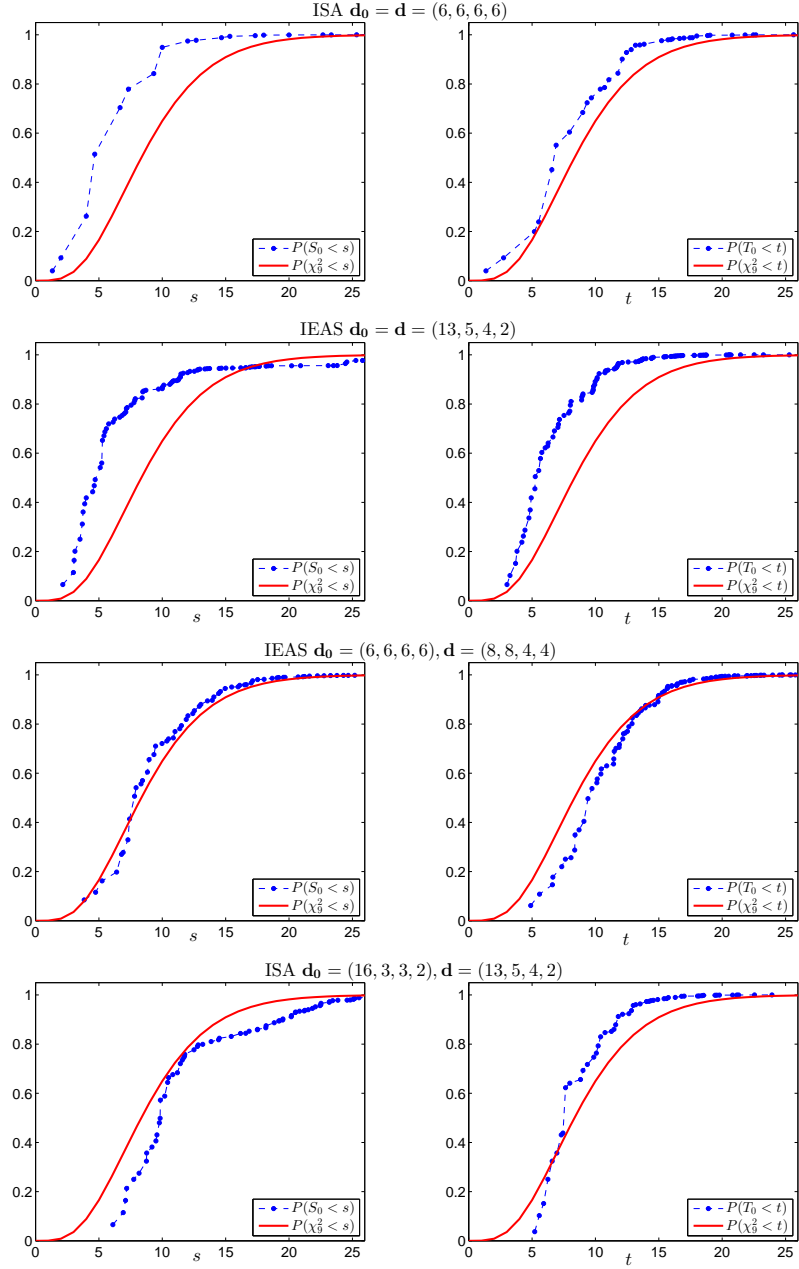


Figure 18: Distributions of S_0 , T_0 and χ_9^2 for some $\text{RSM}(\mathbf{d})$ models and $\text{IEAS}(\mathbf{d}_0)$ or $\text{ISA}(\mathbf{d}_0/2m)$ hypotheses with $n = 4$ and $m = 12$.

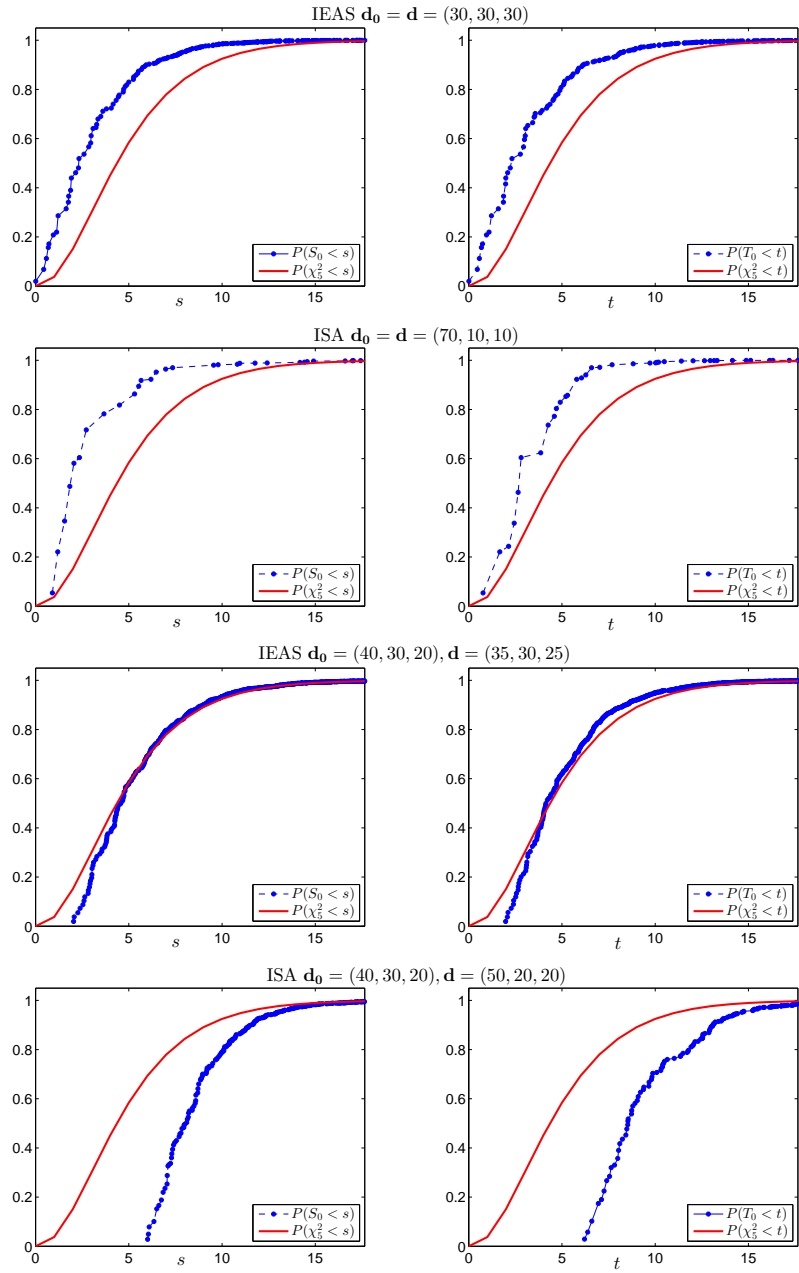


Figure 19: Distributions of S_0 , T_0 and χ_5^2 for some $\text{RSM}(\mathbf{d})$ models and $\text{IEAS}(\mathbf{d}_0)$ or $\text{ISA}(\mathbf{d}_0/2m)$ hypotheses with $n = 3$ and $m = 45$.

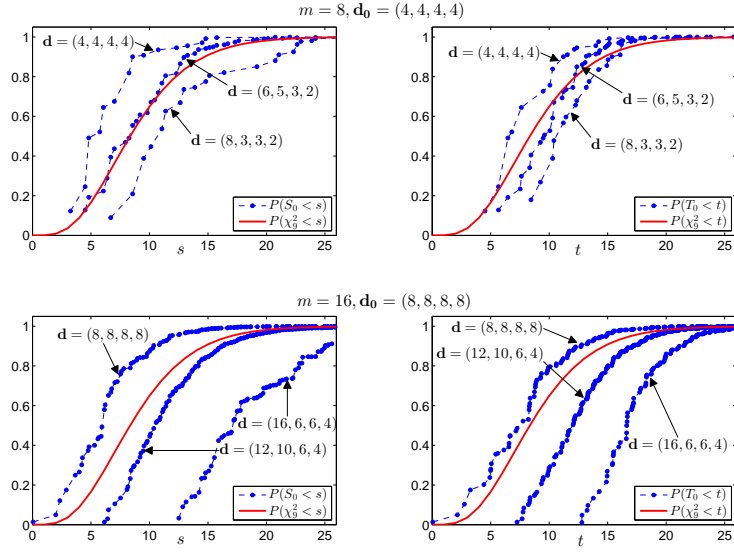


Figure 20: Non-null distributions of S_0 and T_0 for some RSM(\mathbf{d}) models and IEAS(\mathbf{d}_0) hypotheses with flat \mathbf{d}_0 and different \mathbf{d} when m increases.

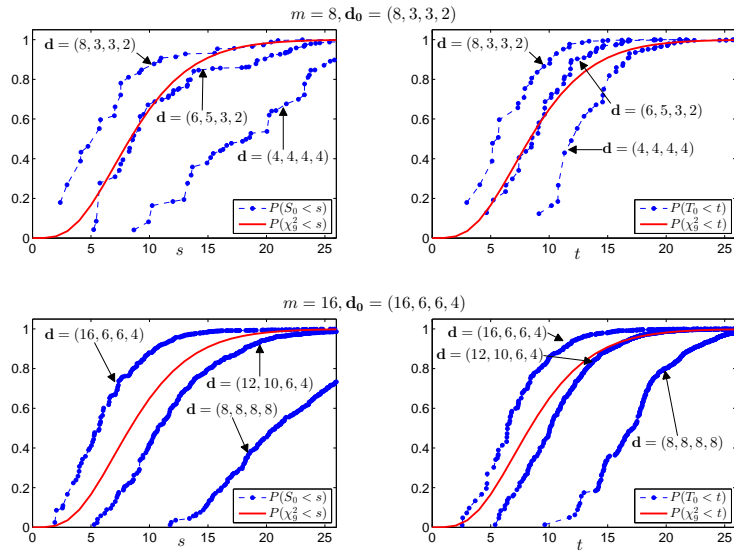


Figure 21: Non-null distributions of S_0 and T_0 for some RSM(\mathbf{d}) models and IEAS(\mathbf{d}_0) hypotheses with skew \mathbf{d}_0 and different \mathbf{d} when m increases.

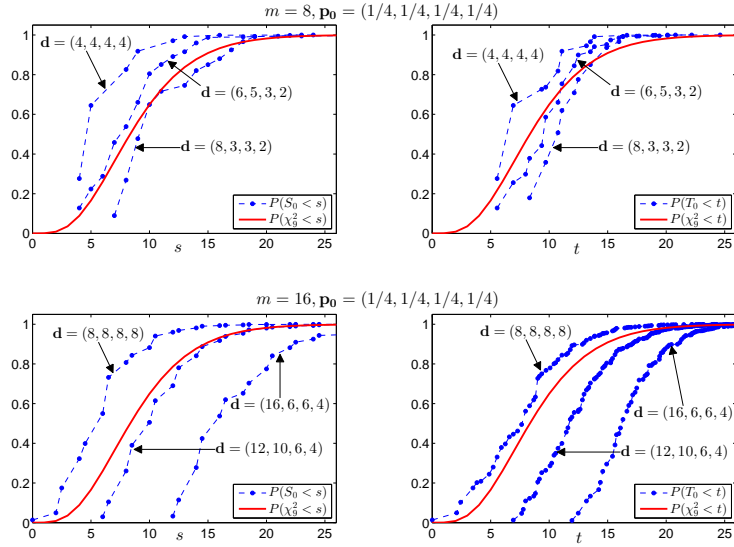


Figure 22: Non-null distributions of S_0 and T_0 for some $\text{RSM}(\mathbf{d})$ models and $\text{ISA}(\mathbf{p}_0)$ hypotheses with flat \mathbf{p}_0 and different \mathbf{d} when m increases.

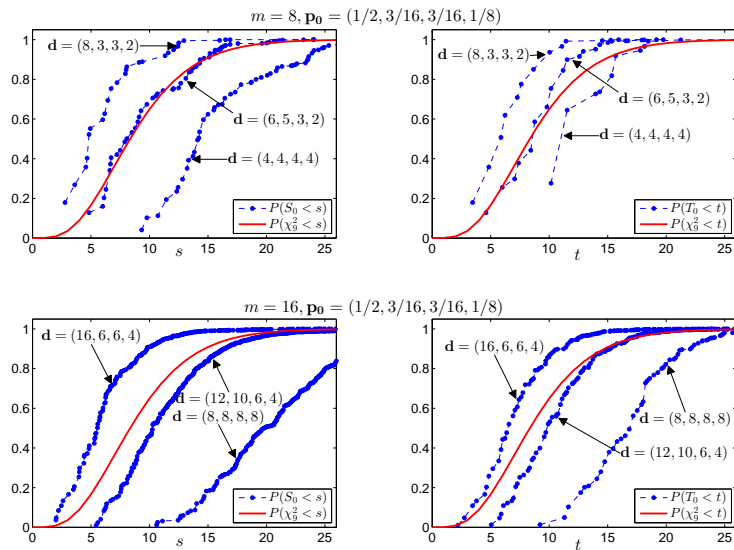


Figure 23: Non-null distributions of S_0 and T_0 for some $\text{RSM}(\mathbf{d})$ models and $\text{ISA}(\mathbf{p}_0)$ hypotheses with skew \mathbf{p}_0 and different \mathbf{d} when m increases.

5 Statistical Tests of a Composite Multigraph Hypothesis

5.1 Test Statistics

The composite multigraph hypothesis might be ISA for unknown \mathbf{p} or IEAS for unknown \mathbf{d} . The parameters have to be estimated from data \mathbf{m} . These estimates are denoted $\hat{\mathbf{p}} = \hat{\mathbf{p}}(\mathbf{m})$ and $\hat{\mathbf{d}} = \hat{\mathbf{d}}(\mathbf{m})$, and they are related according to

$$\hat{\mathbf{p}} = \frac{\hat{\mathbf{d}}}{2m} ,$$

where

$$\hat{d}_i = \sum_{j=1}^n (m_{ij} + m_{ji}) = m_{i\cdot} + m_{\cdot i} \quad \text{for } i = 1, \dots, n ,$$

and $m_{ij} = 0$ for $i > j$. Thus, we have estimated sequences $\hat{\mathbf{Q}} = (\hat{Q}_{ij} : (i, j) \in R)$ in the two cases with composite ISA and IEAS hypotheses. Note that for ISA

$$\hat{Q}_{ij} = \begin{cases} \hat{p}_i^2 & \text{for } i = j \\ 2\hat{p}_i\hat{p}_j & \text{for } i < j , \end{cases}$$

and for IEAS

$$\hat{Q}_{ij} = \begin{cases} \binom{\hat{d}_i}{2} / \binom{2m}{2} & \text{for } i = j \\ \hat{d}_i\hat{d}_j / \binom{2m}{2} & \text{for } i < j . \end{cases}$$

The Pearson goodness-of-fit and divergence statistics are here given as

$$\hat{S} = \sum_{i \leq j} \sum \frac{(m_{ij} - m\hat{Q}_{ij})^2}{m\hat{Q}_{ij}} = \sum_{i \leq j} \sum \frac{m_{ij}^2}{m\hat{Q}_{ij}} - m ,$$

and

$$\hat{D} = \sum_{i \leq j} \sum \frac{m_{ij}}{m} \log \frac{m_{ij}}{m\hat{Q}_{ij}} .$$

Here, \hat{S} and

$$\hat{T} = \frac{2m}{\log e} \hat{D}$$

are asymptotically $\chi^2_{\binom{n}{2}}$ -distributed when the correct model is tested. Note that the number of degrees of freedom here is given as the difference in numbers of estimated free parameters without and with the hypothesis, i.e. $df = (r - 1) - (n - 1) = r - n = \binom{n}{2}$. The critical

regions for these tests are given by values of \hat{S} and \hat{T} above a critical value cv which can be chosen as

$$cv = df + 2\sqrt{2df} = \binom{n}{2} + \sqrt{8\binom{n}{2}}$$

to get a significance level close to 5% given by

$$\alpha = P(\chi_{\binom{n}{2}}^2 > cv) .$$

The power functions $P(\hat{S} > cv)$ and $P(\hat{T} > cv)$ are functions of \mathbf{p} or \mathbf{d} depending on whether an ISA(\mathbf{p}) or IEAS(\mathbf{d}) model is considered. The error probabilities of false rejection and false acceptance are denoted by α and β indexed by \hat{S} and \hat{T} .

Similar to the test statistic approximations described in Section 4.1, S' and S'' are here given by S^* for k chosen as the integer part of μ and $r - n$, respectively. These approximations can be used as alternative test statistics provided the expected values of \hat{S} and \hat{T} are known. Formal expressions for the expected values are complicated to obtain due to that \mathbf{m} is involved also via $\hat{\mathbf{Q}}$ that depends on $\hat{\mathbf{d}}$ which is determined by \mathbf{m} . However, for our theoretical investigation we use complete enumerations of all outcomes of \mathbf{m} and find the expected values and variances numerically. Under an RSM(\mathbf{d}) model the estimated $\hat{\mathbf{d}}$ is always (for any data \mathbf{m}) equal to the \mathbf{d} specified in the model which implies that

$$E(\hat{S}) = E(S_0) = \frac{(m-1)n(n-1)}{2m-3} ,$$

as shown in Section 4.1. The preferences between approximations to the test statistics under IEA models are determined by comparing variances, as mentioned in Section 4.1.

5.2 Test Illustrations for IEAS Models

Consider composite IEAS hypotheses against IEAS(\mathbf{d}) models for multigraphs with 4 vertices and 10 edges. Here, the composite hypotheses include the correct model and the probabilities of false rejection according to \hat{S} and \hat{T} are given in Table 9. For flat \mathbf{d} , both $\alpha_{\hat{S}}$ and $\alpha_{\hat{T}}$ are close or equal to $\alpha = 0.04$ and for skew \mathbf{d} , $\alpha_{\hat{S}}$ remains close or equal to α while $\alpha_{\hat{T}}$ is below. If the composite ISA hypothesis is instead tested against the IEAS(\mathbf{d}) model, the powers of \hat{S} and \hat{T} are almost equal to the values of $\alpha_{\hat{S}}$ and $\alpha_{\hat{T}}$ in Table 9. Thus, both statistics have very poor powers of detecting differences between composite ISA and IEAS hypotheses. Figure 24 illustrates the fit of the distributions of \hat{S} and \hat{T} to that of χ_6^2 . For skew \mathbf{d} there are larger deviations from χ_6^2 for both \hat{S} and \hat{T} than there are for flat \mathbf{d} .

Figures 25 and 26 shows the null and non-null distributions of \hat{S} and \hat{T} for some IEAS(\mathbf{d}) models with flat and skew \mathbf{d} when m increases as multiples of the specified \mathbf{d} . The null distributions correspond to composite IEAS hypotheses while non-null distributions correspond to composite ISA hypotheses. The convergence of the null distributions for flat \mathbf{d} is

rapid towards the asymptotic distribution, especially for \hat{S} . However, the convergence of the null distributions for skew \mathbf{d} is slower for both statistics. The non-null distributions of both statistics also seem to converge to the asymptotic null distributions. Thus, for small and large m , it is difficult to detect differences between composite ISA and IEAS hypotheses.

The expected values and variances of \hat{S} and \hat{T} , and of their approximations \hat{S}' , \hat{S}'' , \hat{T}' and \hat{T}'' are presented in Table 10, where the versions that are not preferred are shaded. For flat \mathbf{d} , the variances of \hat{S} are roughly twice their expected values, which are approximately equal to 6. This indicates a good fit to the χ_6^2 -distribution in terms of the first two moments. This is also noted by \hat{S}'' being preferred to \hat{S}' . Further, for flat \mathbf{d} , \hat{S}'' and \hat{T}' are preferred, while for the majority of cases with skew \mathbf{d} , \hat{T}'' and \hat{S}' are preferred. Two particular cases are $\mathbf{d}=(6, 6, 6, 2)$ and $\mathbf{d}=(6, 5, 5, 4)$ where the variances of the approximations are equal so that any one of them can be preferred.

Table 9: Probabilities of false rejection according to \hat{S} and \hat{T} when model is IEAS(\mathbf{d}) and a composite IEAS hypothesis is tested for $n = 4$ and $m = 10$. $\alpha = 0.04$.

\mathbf{d}	(14, 2, 2, 2)	(12, 3, 3, 2)	(9, 7, 2, 2)	(8, 8, 2, 2)	(6, 6, 6, 2)	(6, 5, 5, 4)	(5, 5, 5, 5)
\hat{S}	0.04	0.04	0.03	0.03	0.02	0.03	0.03
\hat{T}	0.00	0.01	0.01	0.01	0.03	0.04	0.04

Table 10: Moments of \hat{S} , \hat{S}' , \hat{S}'' , \hat{T} , \hat{T}' and \hat{T}'' when model is IEAS(\mathbf{d}) and a composite IEAS hypothesis is tested for $n = 4$ and $m = 10$. The unshaded rows correspond to the best approximation to the test statistics.

\mathbf{d}		(14, 2, 2, 2)	(12, 3, 3, 2)	(9, 7, 2, 2)	(8, 8, 2, 2)	(6, 6, 6, 2)	(6, 5, 5, 4)	(5, 5, 5, 5)
\hat{S}	Mean	3.69	4.59	4.56	4.57	5.31	5.88	5.94
	Variance	18.50	16.12	12.79	12.71	10.57	11.32	11.08
\hat{S}'	Mean	3.69	4.59	4.56	4.57	5.31	5.88	5.94
	Variance	9.09	10.52	10.39	10.42	11.29	13.85	14.11
\hat{S}''	Mean	3.69	4.59	4.56	4.57	5.31	5.88	5.94
	Variance	4.54	7.02	6.92	6.95	9.41	11.54	11.76
\hat{T}	Mean	3.30	4.57	5.04	5.07	6.23	6.96	7.07
	Variance	6.71	7.43	8.29	8.43	8.97	9.20	9.22
\hat{T}'	Mean	3.30	4.57	5.04	5.07	6.23	6.96	7.07
	Variance	7.26	10.45	10.15	10.29	12.95	16.16	14.28
\hat{T}''	Mean	3.30	4.57	5.04	5.07	6.23	6.96	7.07
	Variance	3.63	6.96	8.46	8.57	12.95	16.16	16.66

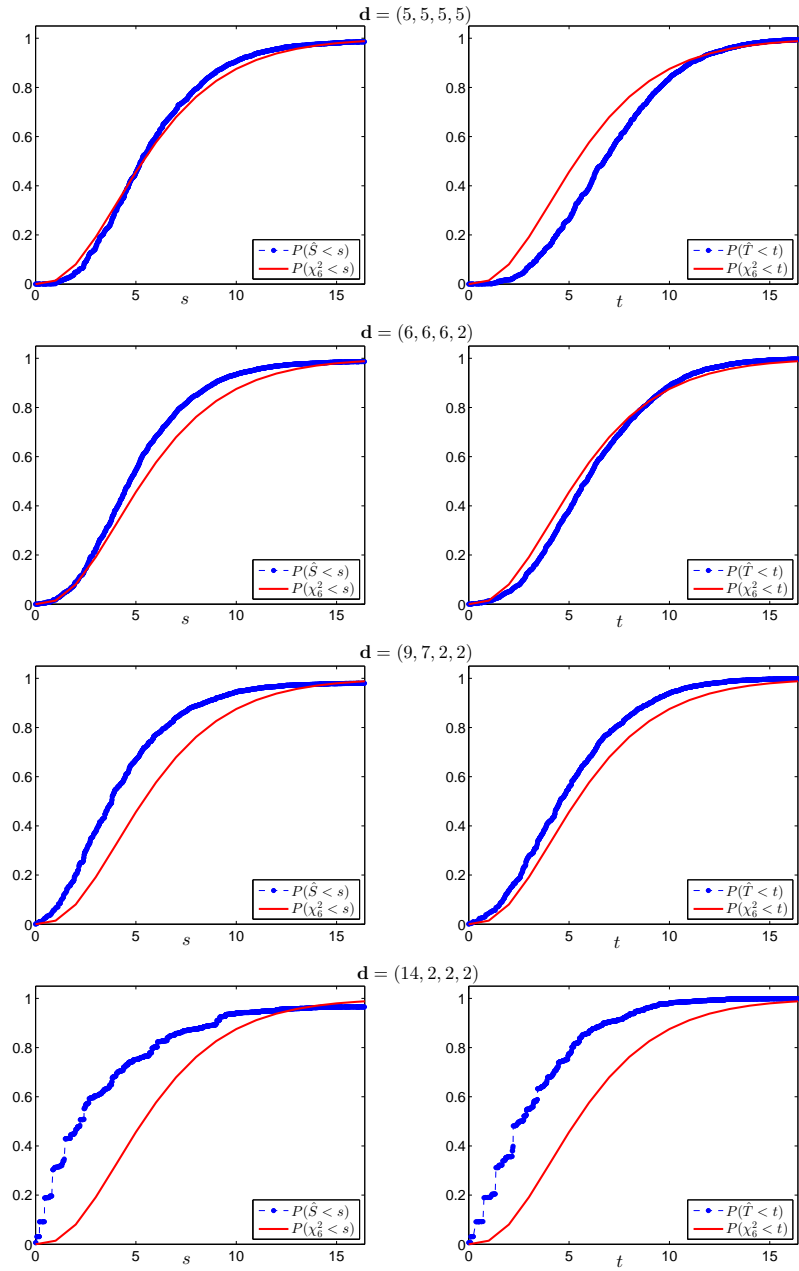


Figure 24: Distributions of \hat{S} , \hat{T} and χ_6^2 for some IEAS(\mathbf{d}) models and composite IEAS hypothesis with $n = 4$ and $m = 10$.

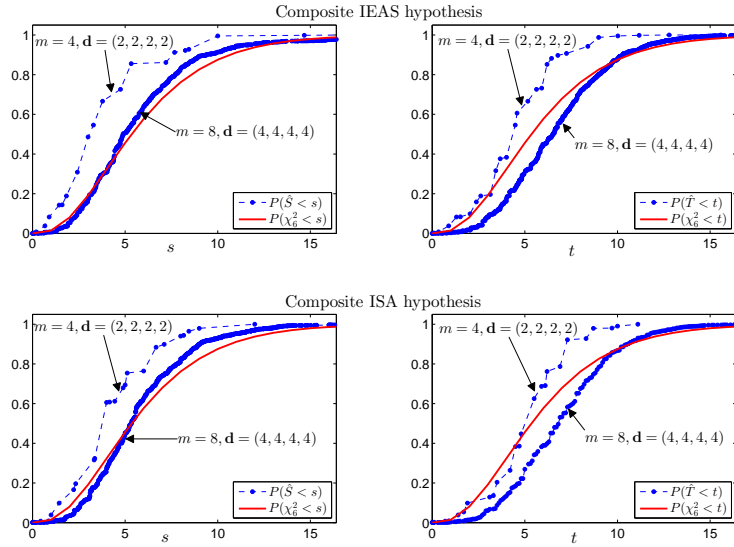


Figure 25: Null and non-null distributions of \hat{S} and \hat{T} for some IEAS(\mathbf{d}) models with flat \mathbf{d} and composite IEAS and ISA hypotheses when m increases.

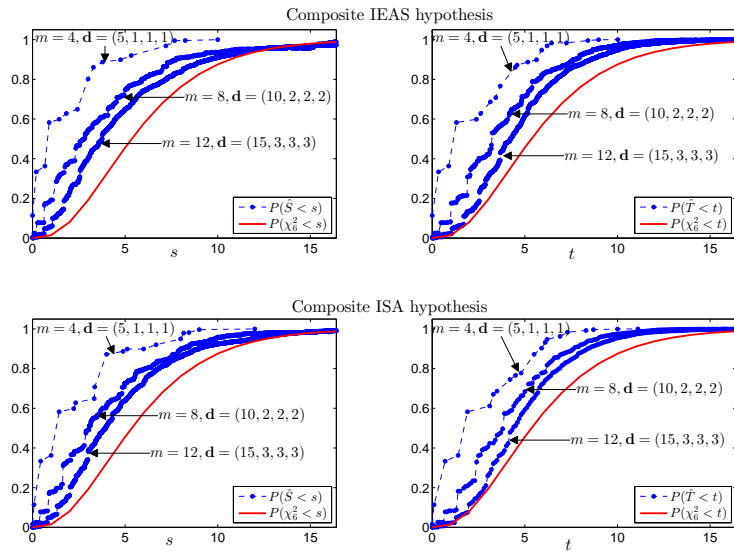


Figure 26: Null and non-null distributions of \hat{S} and \hat{T} for some IEAS(\mathbf{d}) models with skew \mathbf{d} and composite IEAS and ISA hypotheses when m increases.

5.3 Test Illustrations for ISA Models

We now turn to composite ISA hypotheses against $\text{ISA}(\mathbf{p})$ models, and consider tests of multigraphs with 4 vertices and 10 edges. The probabilities of false rejection according to \hat{S} and \hat{T} are given in Table 11 where similar results are seen as for IEAS models in Table 9. For skew \mathbf{p} , $\alpha_{\hat{T}}$ is much below $\alpha = 0.04$, while $\alpha_{\hat{S}}$ is always close to α . If a composite hypothesis IEAS instead of composite ISA is tested against the $\text{ISA}(\mathbf{p})$ model, the powers of \hat{S} and \hat{T} are approximately equal to $\alpha_{\hat{S}}$ and $\alpha_{\hat{T}}$ given in Table 11. As noted before, these poor powers are due to the resemblances between ISA and IEAS models.

Some selected cases from Table 11 are illustrated in Figure 27 where the cumulative distribution functions of \hat{S} and \hat{T} are given. We see that we have fairly good fit between the distributions of both statistics and that of χ_6^2 , except for the very skew $\mathbf{p}=(7/10, 1/10, 1/10, 1/10)$.

In Figures 28 and 29 we illustrate the effect of increasing m on the null and non-null distributions of \hat{S} and \hat{T} for some $\text{ISA}(\mathbf{p})$ models with flat and skew \mathbf{p} . Here, the resemblance between IEAS and ISA models gives similar results as those for composite hypotheses against IEAS models shown in Section 5.2. The convergence of the null distributions is faster for flat \mathbf{p} than for skew \mathbf{p} , and the detection of a composite hypothesis not including the correct model is more difficult as m is increased.

The expected values and variances of all the versions of the test statistics are presented in Table 12 where the unshaded rows correspond to the best approximations. For the majority of cases \hat{S}'' and \hat{T}'' are preferred, except for very skew \mathbf{p} where \hat{S}' and \hat{T}' are preferred.

Table 11: Probabilities of false rejection according to \hat{S} and \hat{T} when model is $\text{ISA}(\mathbf{p})$ and a composite ISA hypothesis is tested for $n = 4$ and $m = 10$. $\alpha = 0.04$.

\mathbf{p}	$(\frac{7}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$	$(\frac{3}{5}, \frac{1}{5}, \frac{1}{10}, \frac{1}{10})$	$(\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$	$(\frac{4}{9}, \frac{1}{3}, \frac{1}{9}, \frac{1}{9})$	$(\frac{3}{8}, \frac{3}{8}, \frac{1}{8}, \frac{1}{8})$	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
\hat{S}	0.03	0.03	0.03	0.03	0.03	0.03
\hat{T}	0.01	0.01	0.02	0.02	0.02	0.05

Table 12: Moments of \hat{S} , \hat{S}' , \hat{S}'' , \hat{T} , \hat{T}' and \hat{T}'' when model is ISA(\mathbf{p}) and a composite ISA hypothesis is tested for $n = 4$ and $m = 10$. The unshaded rows correspond to the best approximation to the test statistics.

\mathbf{p}		$(\frac{7}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$	$(\frac{3}{5}, \frac{1}{5}, \frac{1}{10}, \frac{1}{10})$	$(\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$	$(\frac{4}{9}, \frac{1}{3}, \frac{1}{9}, \frac{1}{9})$	$(\frac{3}{8}, \frac{3}{8}, \frac{1}{8}, \frac{1}{8})$	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
\hat{S}	Mean	4.00	4.59	5.42	4.99	5.22	5.96
	Variance	15.50	11.92	9.81	10.19	9.28	8.41
\hat{S}'	Mean	4.00	4.59	5.42	4.99	5.22	5.96
	Variance	8.01	10.52	11.73	12.43	10.91	14.23
\hat{S}''	Mean	4.00	4.59	5.42	4.99	5.22	5.96
	Variance	5.34	7.01	9.78	8.29	9.09	11.86
\hat{T}	Mean	3.70	4.73	6.01	5.56	5.93	7.24
	Variance	7.17	7.48	8.15	8.72	9.00	9.21
\hat{T}'	Mean	3.70	4.73	6.01	5.56	5.93	7.24
	Variance	9.14	11.18	12.05	12.38	14.06	14.99
\hat{T}''	Mean	3.70	4.73	6.01	5.56	5.93	7.24
	Variance	4.57	7.45	12.05	10.32	11.72	17.49

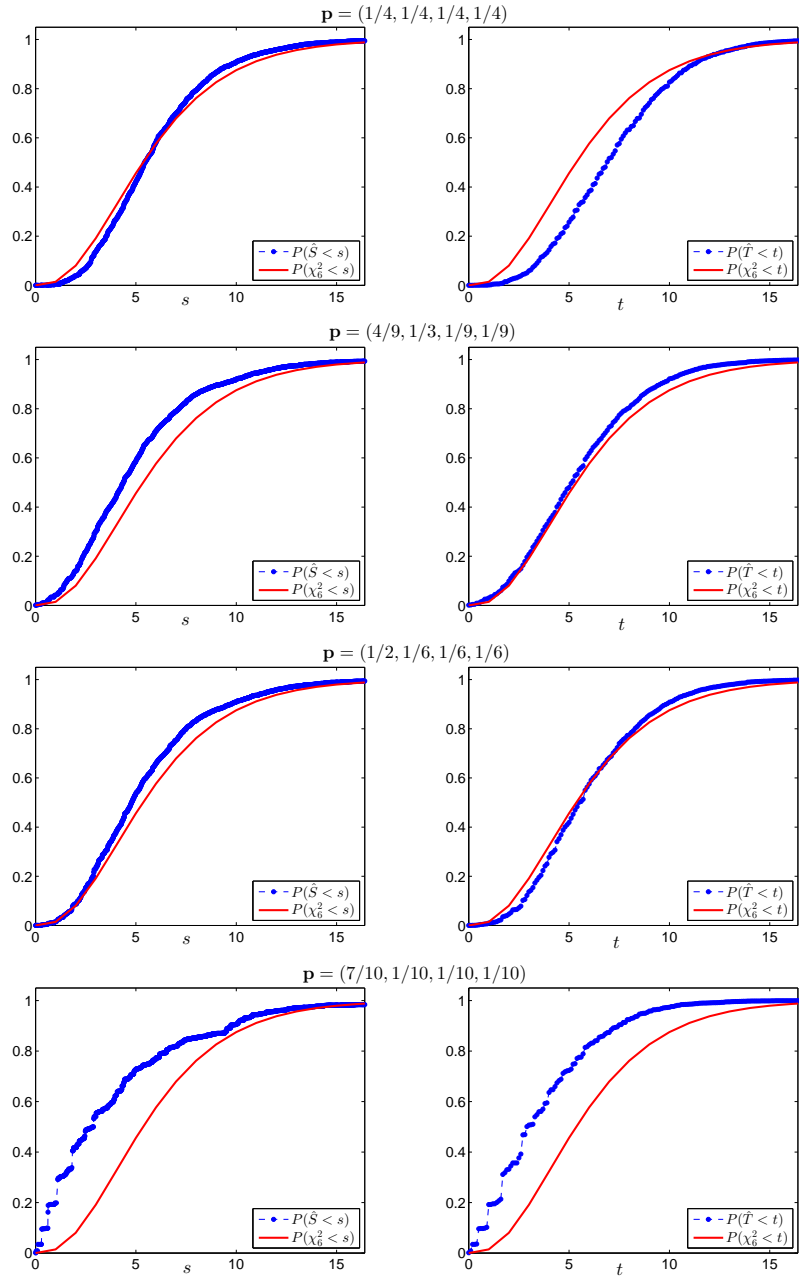


Figure 27: Distributions of \hat{S} , \hat{T} and χ_6^2 for some $\text{ISA}(\mathbf{p})$ models and composite ISA hypothesis with $n = 4$ and $m = 10$.

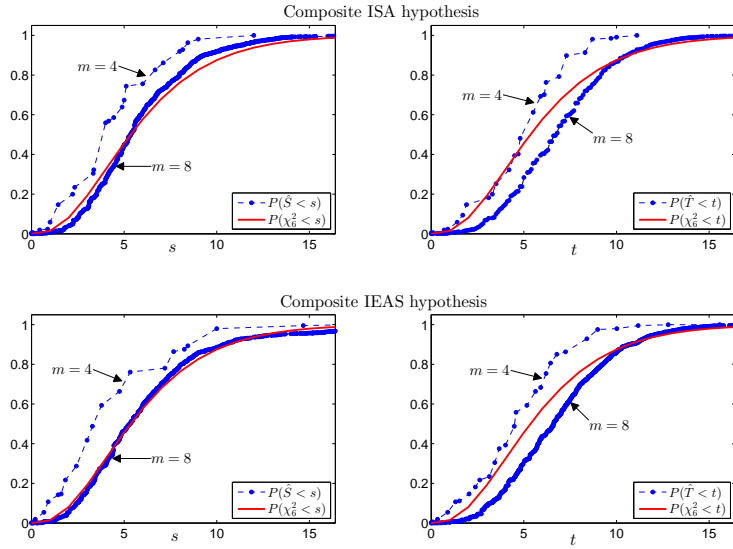


Figure 28: Null and non-null distributions of \hat{S} and \hat{T} for some $\text{ISA}(\mathbf{p})$ models with flat $\mathbf{p}=(1/4, 1/4, 1/4, 1/4)$ and composite ISA and IEAS hypotheses when m increases.

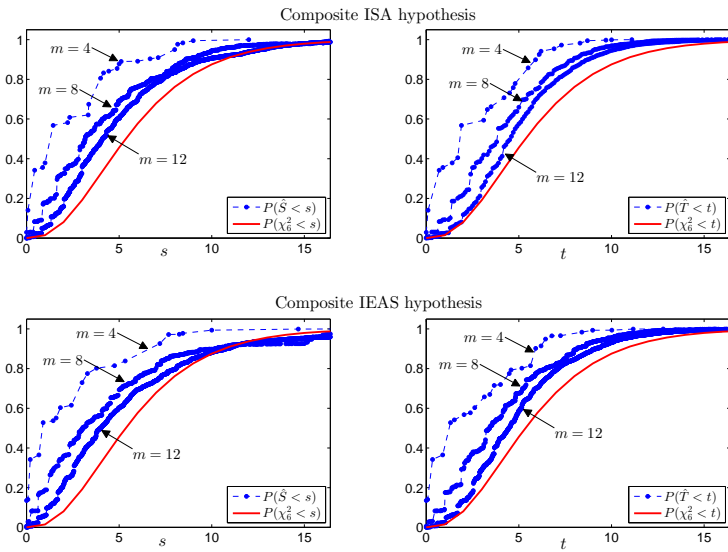


Figure 29: Null and non-null distributions of \hat{S} and \hat{T} for some $\text{ISA}(\mathbf{p})$ models with skew $\mathbf{p}=(5/8, 1/8, 1/8, 1/8)$ and composite ISA and IEAS hypotheses when m increases.

5.4 Test Illustrations for RSM Models

In this section we illustrate some of the consequences of using previously described tests of composite hypotheses against a false IEA model when the true model is RSM. Here, both IEAS and ISA hypotheses are tested against $\text{RSM}(\mathbf{d})$ models. We start by considering multigraphs with 4 vertices and 12 edges. The poor powers according to \hat{S} and \hat{T} of rejecting IEAS and ISA when RSM is true are presented in Table 13. To illustrate the fit of the distributions of the statistics \hat{S} and \hat{T} to that of χ_6^2 , their cumulative distribution functions for some selected cases are shown in Figure 30. For all cases, there is reasonably good fit to χ_6^2 for this rather small m .

The expected values and variances of all the versions of the test statistics are presented in Table 14. For the majority of cases, the variances of \hat{S} are roughly twice their expected values which are equal to 6. This indicates a good fit to the χ_6^2 -distribution in terms of the first two moments. Note that $E(\hat{S})$ under IEAS hypotheses are not dependent on the values in the degree sequence, as mentioned in Section 5.1. For very skew \mathbf{d} , \hat{T}'' is preferred for almost all cases and for the rest of the skew cases and for flat cases, \hat{T}' is preferred.

The poor powers of rejecting IEAS and ISA when RSM is true for multigraphs with 3 vertices and 45 edges are shown in Table 15. We see that $\alpha_{\hat{S}}$ is close to $\alpha = 0.04$ for all cases shown, while \hat{T} is equal, less or greater than α for both skew and flat \mathbf{d} . The fit of the non-null distributions of \hat{S} and \hat{T} to that of χ_3^2 for some selected cases are shown in Figure 31 where we for all cases illustrated see a good fit. The expected values and variances of all the versions of the test statistics are presented in Table 16. For all cases with IEAS hypotheses, and almost all cases with ISA hypotheses, we note a good fit to the χ_3^2 -distribution since the variances of both test statistics are roughly twice their expected values which are equal to 3. This indicates that the approximations are mostly unnecessary for large m .

In Figures 32 and 33 the effects of increasing m on the non-null distributions of \hat{S} and \hat{T} for some $\text{RSM}(\mathbf{d})$ models with flat and skew \mathbf{d} are illustrated. For all cases illustrated we see that these distributions are very close to the asymptotic null distribution. Further, the effect from increasing m on the non-null distributions is small. Thus it can be concluded that no matter the size of m , it is difficult to detect a false composite hypothesis under an RSM model, just as it is difficult to detect a false composite hypothesis under IEA models as demonstrated in Figures 25-26 for IEAS models, and in Figures 28-29 for ISA models.

Table 13: Power according to \hat{S} and \hat{T} when model is RSM(\mathbf{d}) and a composite IEAS or ISA hypothesis is tested for $n = 4$ and $m = 12$. $\alpha = 0.04$.

		\mathbf{d}	(18, 2, 2, 2)	(16, 3, 3, 2)	(13, 5, 4, 2)	(8, 8, 4, 4)	(7, 7, 7, 3)	(7, 6, 6, 5)	(6, 6, 6, 6)
IEAS	\hat{S}		0.14	0.09	0.06	0.03	0.04	0.04	0.03
	\hat{T}		0.02	0.02	0.03	0.05	0.06	0.08	0.07
ISA	\hat{S}		0.04	0.08	0.06	0.03	0.03	0.03	0.02
	\hat{T}		0.02	0.01	0.02	0.06	0.06	0.09	0.06

Table 14: Moments of \hat{S} , \hat{S}' , \hat{S}'' , \hat{T} , \hat{T}' and \hat{T}'' when model is RSM(\mathbf{d}) and a composite IEAS or ISA hypothesis is tested for $n = 4$ and $m = 12$. The unshaded rows correspond to the best approximations to the test statistics.

		\mathbf{d}	Composite IEAS hypothesis						
			(18, 2, 2, 2)	(16, 3, 3, 2)	(13, 5, 4, 2)	(8, 8, 4, 4)	(7, 7, 7, 3)	(7, 6, 6, 5)	(6, 6, 6, 6)
\hat{S}	Mean		6.29	6.29	6.29	6.29	6.29	6.29	6.29
	Variance		66.26	32.41	26.50	10.18	11.83	9.64	9.63
\hat{S}'	Mean		6.29	6.29	6.29	6.29	6.29	6.29	6.29
	Variance		13.18	13.17	13.17	13.17	13.17	13.17	13.17
\hat{S}''	Mean		6.29	6.29	6.29	6.29	6.29	6.29	6.29
	Variance		13.18	13.17	13.17	13.17	13.17	13.17	13.17
\hat{T}	Mean		3.80	5.08	6.26	7.34	7.39	7.83	7.90
	Variance		9.06	7.25	7.57	11.43	11.77	11.54	11.56
\hat{T}'	Mean		3.80	5.08	6.26	7.34	7.39	7.83	7.90
	Variance		9.63	10.32	13.04	15.40	15.62	17.51	17.82
\hat{T}''	Mean		3.80	5.08	6.26	7.34	7.39	7.83	7.90
	Variance		4.81	8.60	13.04	17.97	18.23	20.43	20.80
		\mathbf{d}	Composite ISA hypothesis						
			(18, 2, 2, 2)	(16, 3, 3, 2)	(13, 5, 4, 2)	(8, 8, 4, 4)	(7, 7, 7, 3)	(7, 6, 6, 5)	(6, 6, 6, 6)
\hat{S}	Mean		5.12	5.52	5.76	6.09	6.07	6.18	6.19
	Variance		23.58	13.19	10.87	8.41	9.07	8.70	8.78
\hat{S}'	Mean		5.12	5.52	5.76	6.09	6.07	6.18	6.19
	Variance		10.50	12.17	13.26	12.36	12.29	12.71	12.76
\hat{S}''	Mean		5.12	5.52	5.76	6.09	6.07	6.18	6.19
	Variance		8.75	10.14	11.05	12.36	12.29	12.71	12.76
\hat{T}	Mean		3.86	5.16	6.33	7.42	7.46	7.90	7.98
	Variance		6.78	5.36	6.40	10.83	11.44	11.33	11.37
\hat{T}'	Mean		3.86	5.16	6.33	7.42	7.46	7.90	7.98
	Variance		9.92	10.66	13.34	15.73	15.91	17.82	18.18
\hat{T}''	Mean		3.86	5.16	6.33	7.42	7.46	7.90	7.98
	Variance		4.96	8.88	13.34	18.35	18.56	20.79	21.21

Table 15: Power according to \hat{S} and \hat{T} when model is RSM(\mathbf{d}) and a composite IEAS or ISA hypothesis is tested for $n = 3$ and $m = 45$. $\alpha = 0.04$.

		\mathbf{d}	(70, 10, 10)	(65, 15, 10)	(50, 20, 20)	(45, 35, 10)	(40, 30, 20)	(35, 30, 25)	(30, 30, 30)
IEAS	\hat{S}		0.04	0.04	0.04	0.04	0.05	0.05	0.05
	\hat{T}		0.02	0.04	0.07	0.05	0.06	0.06	0.07
ISA	\hat{S}		0.03	0.04	0.04	0.04	0.04	0.05	0.05
	\hat{T}		0.02	0.04	0.06	0.05	0.06	0.06	0.06

Table 16: Moments of \hat{S} , \hat{S}' , \hat{S}'' , \hat{T} , \hat{T}' and \hat{T}'' when model is RSM(\mathbf{d}) and a composite IEAS or ISA hypothesis is tested for $n = 3$ and $m = 45$. The unshaded rows correspond to the best approximations to the test statistics.

		\mathbf{d}	(70, 10, 10)	(65, 15, 10)	Composite IEAS hypothesis				
					(50, 20, 20)	(45, 35, 10)	(40, 30, 20)	(35, 30, 25)	(30, 30, 30)
\hat{S}	Mean		3.03	3.03	3.03	3.03	3.03	3.03	3.03
	Variance		6.62	6.12	5.71	6.11	5.78	5.81	5.83
\hat{S}'	Mean		3.03	3.03	3.03	3.03	3.03	3.03	3.03
	Variance		6.14	6.14	6.14	6.14	6.14	6.14	6.14
\hat{S}''	Mean		3.03	3.03	3.03	3.03	3.03	3.03	3.03
	Variance		6.14	6.14	6.14	6.14	6.14	6.14	6.14
\hat{T}	Mean		3.43	3.48	3.31	3.22	3.21	3.15	3.14
	Variance		3.66	5.28	7.35	5.48	6.94	6.77	6.66
\hat{T}'	Mean		3.43	3.48	3.31	3.22	3.21	3.15	3.14
	Variance		7.82	8.09	7.30	6.92	6.88	6.63	6.57
\hat{T}''	Mean		3.43	3.48	3.31	3.22	3.21	3.15	3.14
	Variance		7.82	8.09	7.30	6.92	6.88	6.63	6.57

		\mathbf{d}	(70, 10, 10)	(65, 15, 10)	Composite ISA hypothesis				
					(50, 20, 20)	(45, 35, 10)	(40, 30, 20)	(35, 30, 25)	(30, 30, 30)
\hat{S}	Mean		2.91	2.95	3.01	2.98	3.03	3.03	3.03
	Variance		5.55	5.36	5.50	5.60	5.66	5.75	5.78
\hat{S}'	Mean		2.91	2.95	3.01	2.98	3.03	3.03	3.03
	Variance		8.49	8.69	6.05	8.88	6.10	6.13	6.14
\hat{S}''	Mean		2.91	2.95	3.01	2.98	3.03	3.03	3.03
	Variance		5.66	5.79	6.05	5.92	6.10	6.13	6.14
\hat{T}	Mean		3.45	3.50	3.32	3.23	3.23	3.17	3.15
	Variance		3.47	5.23	7.45	5.42	7.02	6.82	6.73
\hat{T}'	Mean		3.45	3.50	3.32	3.23	3.23	3.17	3.15
	Variance		7.92	8.16	7.35	6.98	6.95	6.71	6.60
\hat{T}''	Mean		3.45	3.50	3.32	3.23	3.23	3.17	3.15
	Variance		7.92	8.16	7.35	6.98	6.95	6.71	6.60

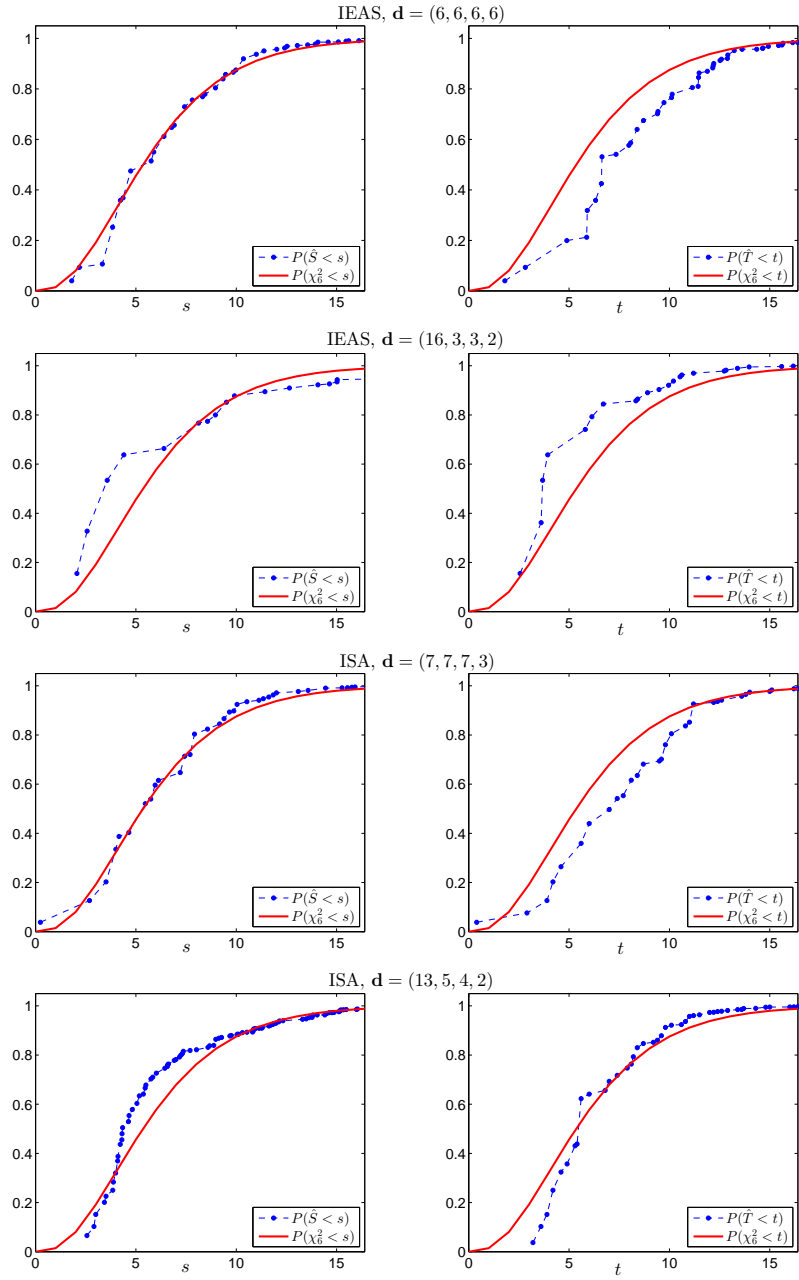


Figure 30: Distributions of \hat{S} , \hat{T} and χ_6^2 for some RSM(\mathbf{d}) models and composite IEAS or ISA hypotheses with $n = 4$ and $m = 12$.

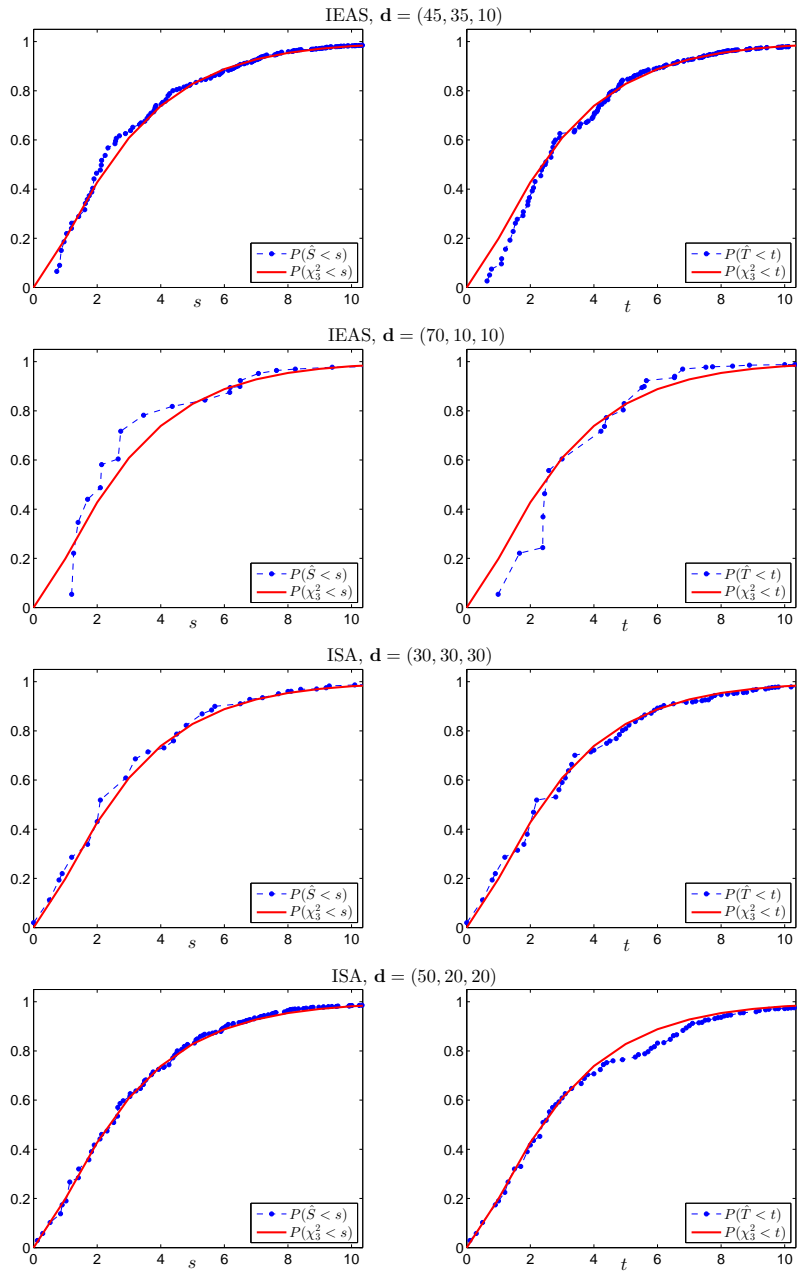


Figure 31: Distributions of \hat{S} , \hat{T} and χ_6^2 for some RSM(\mathbf{d}) models and composite IEAS or ISA hypotheses with $n = 3$ and $m = 45$.

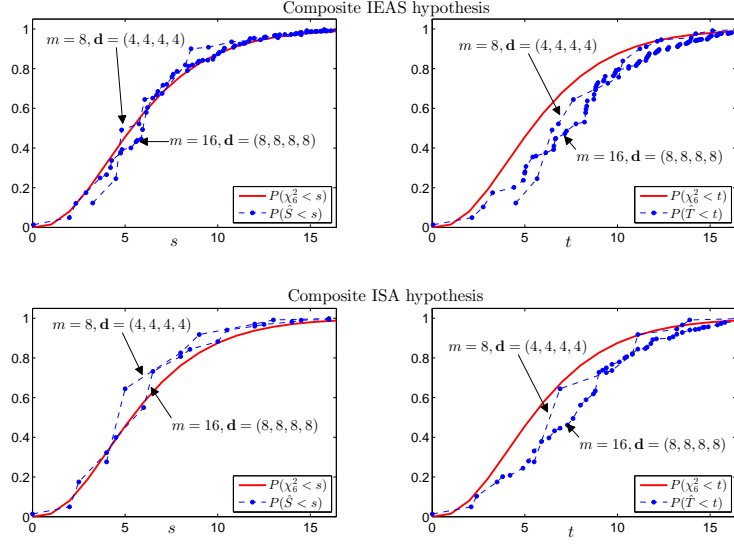


Figure 32: Non-null distributions of \hat{S} and \hat{T} for some RSM(\mathbf{d}) models with flat \mathbf{d} and composite IEAS and ISA hypotheses when m increases.

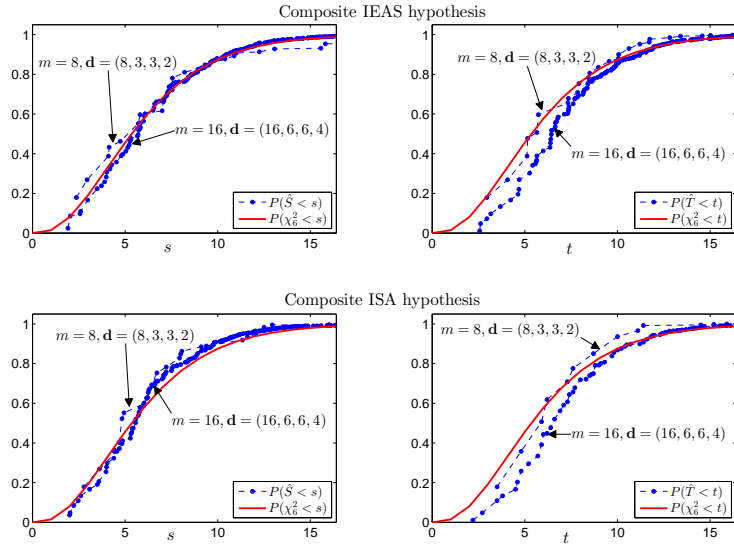


Figure 33: Non-null distributions of \hat{S} and \hat{T} for some RSM(\mathbf{d}) models with skew \mathbf{d} and composite IEAS and ISA hypotheses when m increases.

References

Andersen, E.B. (1980), *Discrete Statistical Models with Social Science Applications*, Amsterdam: North-Holland.

Cox, D.R. and Hinkley, D.V. (1974), *Theoretical Statistics*, London: Chapman & Hall.

Frank, O. (2011), Statistical Information Tools for Multivariate Discrete Data, in *Modern Mathematical Tools and Techniques in Capturing Complexity*, eds. L. Pardo, N. Balakrishnan and M. Ángeles Gil, Berlin: Springer Verlag, 177–190.

Frank, O. and Shafie, T. (2012), Complexity of Families of Multigraphs, to appear in *JSM Proceedings*, Section on Statistical Graphics, Alexandria, VA: American Statistical Association.

Kullback, S. (1959), *Information Theory and Statistics*, New York: Wiley.

Shafie, T. (2012), Random Stub Matching Models of Multigraphs, *Research Report 2012:1*, Department of Statistics, Stockholm University.

Some Multigraph Algorithms

Termeh Shafie

Abstract

Several algorithms for generating and analyzing multigraphs under two different multigraph models are presented including the following: to find distributions of complexity measures in different random multigraphs, to analyze the local and global structure of multigraphs under different multigraph models using information theoretic tools based on entropy, and to test simple or composite hypotheses concerning random multigraphs.

Keywords: multigraph, algorithm, graph enumeration, complexity, multiplicity, entropy, information divergence, goodness-of-fit.

Introduction

There are many graph theoretic algorithms available in the literature. In this article, multigraph algorithms used in the articles by Frank and Shafie (2012), Shafie (2012a) and Shafie (2012b) are presented. Algorithms are given for generating multigraphs under two different multigraph models. The first 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. The second multigraph model is obtained by independent edge assignments (IEA) according to a common probability distribution over the sites. Two different methods for obtaining an approximate IEA model from an RSM model are also considered. The first method is obtained by assuming that the stubs are randomly generated and independently assigned to vertices, called independent stub assignments (ISA) and the second method of obtaining an approximate IEA model is to ignore the dependency between edges in the RSM model and assume independent edge assignments of stubs (IEAS).

Algorithms are also given for analyzing and testing different multigraph models using information theoretic tools based on entropy. In particular, algorithms are given for using the local and global distributions under RSM and IEA to calculate moments and entropies,

Department of Statistics, Stockholm University, S-106 91 Stockholm, termeh.shafie@stat.su.se

and for comparisons between distributions by information divergence. Further, special algorithms are developed for analyzing the complexity of multigraphs which is defined and quantified by the distribution of edge multiplicities.

All algorithms are developed by me and presented in a brief algorithmic style. They have been used for research work presented in the references, and details about concepts and notations used in the algorithms can be found there. Note that there might be more efficient alternatives available in the computer science literature. Such efficiency might be required in order to apply the methods to large multigraphs or to extensive calculations with many multigraphs. Since this has not been needed during the methodological development, no attempts have been made to get optimal algorithms.

List of Algorithms

1	Generating non-decreasing edge sequences for multigraphs with fixed number of vertices and edges	3
2	Complexity of multigraphs with fixed number of vertices and edges	4
3	Generating non-decreasing permutations of a stub multiplicity sequence	4
4	Complexity distribution and multigraph distribution for multigraphs under RSM	5
5	Entropies of trivariate edge multiplicity distributions under RSM and IEA	6
6	Entropies and moments of marginal loop distributions under RSM and IEA	7
7	Entropies and moments of marginal non-loop distributions under RSM and IEA	8
8	Information divergence between trivariate edge multiplicity distributions under RSM and IEA	9
9	Information divergence between marginal loop multiplicity distributions under RSM and IEA	9
10	Information divergence between marginal non-loop multiplicity distributions under RSM and IEA	9
11	Entropies of and information divergence between distributions of multigraphs under RSM and IEA	10
12	Entropy approximations of distributions of multigraphs under RSM	11
13	Approximations of the probability that an RSM multigraph is simple	12
14	Distribution of the multiplicity sequence under an RSM, IEAS or ISA model	13
15	Statistical tests of a simple IEA hypothesis	14
16	Statistical tests of a composite IEAS hypothesis	15
17	Statistical tests of a composite ISA hypothesis	16

Algorithm 1: Generating non-decreasing edge sequences for multigraphs with fixed number of vertices and edges

Input: Number of vertices n and number of edges m

Output: A list \mathbf{Z} with all possible non-decreasing edge sequences

```
1  $r \leftarrow \binom{n+1}{2}$ 
2  $t \leftarrow 1$ 
3 for  $i \leftarrow 1$  to  $m$  do
4    $e(t, i) \leftarrow 1$ 
5  $k \leftarrow m$ 
6 while  $k > 0$  do
7   while  $e(t, k) < r$  do
8      $t \leftarrow t + 1$ 
9     for  $i \leftarrow 1$  to  $m$  do
10      if  $i < k$  then
11         $e(t, i) \leftarrow e(t-1, i)$ 
12      else if  $i \geq k$  and  $e(t, i) + 1 \leq r$  then
13         $e(t, i) \leftarrow e(t-1, k) + 1$ 
14       $k \leftarrow m$ 
15    if  $e(t, k) = r$  then
16       $k \leftarrow k - 1$ 
17  $\mathbf{Z} \leftarrow e$ 
18 for  $i \leftarrow 1$  to  $n$  do
19   for  $j \leftarrow 1$  to  $n$  do
20     if  $i \leq j$  then
21        $A(i, j) = 1$ 
22     else
23        $A(i, j) = 0$ 
24 foreach row in  $\mathbf{Z}$  do
25   foreach column  $c \leftarrow 1$  to  $m$  do
26     for  $i \leftarrow 1$  to  $n$  do
27       for  $j \leftarrow 1$  to  $n$  do
28         if  $A(i, j) > 0$  then
29            $a \leftarrow$  row index  $i$ 
30            $b \leftarrow$  column index  $j$ 
31           recode  $c \leftarrow (a, b)$ 
32 return  $\mathbf{Z}$ 
```

Algorithm 2: Complexity of multigraphs with fixed number of vertices and edges

Input: Number of vertices n and number of edges m , edge loops *allowed* or *forbidden*

Output: Properties of multigraphs including complexity sequences $\mathbf{r} = (r_0, \dots, r_m)$ and summary complexity measure t for each possible non-decreasing edge sequence

```
1 Call Algorithm 1 for  $e$  and  $\mathbf{Z}$ 
2  $r \leftarrow \binom{n+1}{2}$ 
3 foreach row  $i$  in  $\mathbf{Z}$  do
4   for  $j \leftarrow 1$  to  $r$  do
5     Multiplicity sequence  $\mathbf{m}(i, j) \leftarrow$  frequency of  $e(i) = j$ 
6   for  $j \leftarrow 0$  to  $m$  do
7     Complexity sequence  $\mathbf{r}(i, j + 1) \leftarrow$  frequency of  $\mathbf{m}(i) = j$ 
8    $M(i) \leftarrow$  upper triangular matrix containing elements of  $\mathbf{m}(i)$ 
9   Number of edge loops  $m_1(i) \leftarrow$  sum over all diagonal elements  $M(i)$ 
10  if  $m_1(i) > 0$  then
11     $I_{m_1} \leftarrow 1$ 
12 if forbidden then
13   Remove all rows in  $\mathbf{Z}$  where  $I_{m_1} = 1$ 
14   Repeat steps 3-7
15 foreach row  $i$  in  $\mathbf{Z}$  do
16    $mFac(i) \leftarrow$  product over the factorial of each element in  $\mathbf{m}(i)$ 
17   Complexity summary measure  $t(i) \leftarrow m_1(i) + \log_2 mFac(i)$ 
18   for  $u \leftarrow 1$  to  $n$  do
19     Degree sequence  $\mathbf{d}(i) \leftarrow$  frequency of  $\mathbf{Z}(i) = u$ 
20 return List with columns containing  $\mathbf{Z}$ ,  $\mathbf{d}$ ,  $\mathbf{m}$ ,  $m_1$ ,  $\mathbf{r}$ ,  $t$ 
```

Algorithm 3: Generating non-decreasing permutations of a stub multiplicity sequence

Input: Degree or stub multiplicity sequence $\mathbf{d} = (d_1, d_2, \dots, d_n)$

Output: A list \mathbf{S} with all non-decreasing permutations of a stub sequence

```
1 Number of vertices  $n \leftarrow$  number of columns in  $\mathbf{d}$ 
2 Number of edges  $m \leftarrow$  half of the sum of all element values in  $\mathbf{d}$ 
3  $t \leftarrow 1$ 
4  $\mathbf{s}(t) \leftarrow [1^{d_1} 2^{d_2} \dots n^{d_n}]$ 
5  $k \leftarrow m - 1$ 
6 while  $k > 0$  do
7    $W_k(t) \leftarrow [1^{d_1(t,k)} \dots n^{d_n(t,k)}]$ , an ordered sequence of vertices in the edges  $(e_k(t), \dots, e_m(t))$ 
8   Try to re-order  $W_k(t)$  as a non-decreasing sequence of edges  $(f_k(t), \dots, f_m(t))$  above  $e_k(t)$  so that
    $f_k(t) \leftarrow e_k(t) + 1 \leq f_{k+1}(t) \leq \dots \leq f_m(t)$ , where  $e_k(t) + 1$  means that its second vertex is increased by 1
9   if re-order possible then
10     $\mathbf{s}(t + 1) \leftarrow [e_1(t), \dots, e_{k-1}(t), f_k(t), \dots, f_m(t)]$ 
11     $t \leftarrow t + 1$ 
12   else
13     $k \leftarrow k - 1$ 
14  $\mathbf{S} \leftarrow \mathbf{s}$ 
15 return  $\mathbf{S}$ 
```

Algorithm 4: Complexity distribution and multigraph distribution for multigraphs under RSM

Input: Degree or stub multiplicity sequence $\mathbf{d} = (d_1, d_2, \dots, d_n)$
Output: Lists including complexity distributions P_t and multigraph distributions $P_{\mathbf{Z}}$

- 1 Number of vertices $n \leftarrow$ number of columns in \mathbf{d}
- 2 Number of edges $m \leftarrow$ half of the sum of all element values in \mathbf{d}
- 3 Call **Algorithm 3** for $\mathbf{S} \leftarrow$ list of all permutations of the stub sequence
- 4 $\mathbf{Z} \leftarrow \mathbf{S}$
- 5 Follow steps 2-19 in **Algorithm 2** to find $\mathbf{m}, m_1, \mathbf{r}, t$
- 6 **foreach** row i in \mathbf{Z} **do**
- 7 **if** any element in $\mathbf{m}(i) \geq 2$ **then**
- 8 Multiple edge indicator $I_{m_2}(i) \leftarrow 1$
- 9 $x \leftarrow$ a vector containing elements in $\mathbf{m}(i) \geq 2$
- 10 $C \leftarrow$ product over the factorial of each element in x
- 11 **else**
- 12 Multiple edge indicator $I_{m_2}(i) \leftarrow 0$
- 13 **if** $I_{m_2}(i) = 1$ **then**
- 14 Total number of possible shift permutations between edge pairs $SPB(i) \leftarrow m!/C$
- 15 **else**
- 16 Total number of possible shift permutations between edge pairs $SPB(i) \leftarrow m!$
- 17 **if** $m_1(i) = m$ **then**
- 18 Total number of possible shift permutations within edge pairs $SPW(i) \leftarrow 1$
- 19 **else**
- 20 Total number of possible shift permutations within edge pairs $SPW(i) \leftarrow 2^{m-m_1(i)}$
- 21 Total number of permutations of each edge sequence $K_{\mathbf{Z}}(i) \leftarrow SPB(i) \cdot SPW(i)$
- 22 Total number of multigraphs $K_{\mathbf{d}} \leftarrow$ sum over all elements in $K_{\mathbf{Z}}$
- 23 **foreach** row i in \mathbf{Z} **do**
- 24 Probability of multigraph $P_{\mathbf{Z}}(i) \leftarrow K_{\mathbf{Z}}(i)/K_{\mathbf{d}}$
- 25 $t_{UNI} \leftarrow$ unique values of t
- 26 **foreach** row i in t **do**
- 27 Number of complexity value $K_t(i) \leftarrow$ frequency of $t_{UNI} = t(i)$
- 28 Probability of complexity value $P_t(i) \leftarrow$ sum over all $P_{\mathbf{Z}}$ where $t_{UNI} = t(i)$
- 29 **return** List with columns containing $\mathbf{Z}, P_{\mathbf{Z}}, K_{\mathbf{Z}}, \mathbf{m}, m_1, \mathbf{r}$ and list with columns containing t_{UNI}, K_t, P_t

Algorithm 5: Entropies of trivariate edge multiplicity distributions under RSM and IEA

Input: Number of edges m , degrees d_i and d_j at vertices i and j
Output: The entropies h , entropy upper bounds $Maxh$, and entropy approximations $Apxh$ of $P((m_{ii}, m_{jj}, m_{ij}) = (u, v, w))$ under RSM and IEA

```

1  $a \leftarrow d_i$ ,  $b \leftarrow d_j$ , and  $c \leftarrow \min(a, b)$ 
2  $i \leftarrow 0$ 
3 for  $w \leftarrow 0$  to  $c$  do
4   for  $u \leftarrow 0$  to  $\lfloor (a-w)/2 \rfloor$  do
5     for  $v \leftarrow 0$  to  $\lfloor (b-w)/2 \rfloor$  do
6       if  $(m - a - b + u + v + w) \geq 0$  then
7          $i = i + 1$ 
8          $UVW(i) \leftarrow [u \ v \ w]$ 
9          $P(i) \leftarrow \frac{m! \ a! \ b! \ 2^{(a+b-2u-2v-w)} (2m-a-b)!}{u! \ v! \ w! (a-2u-w)! (b-2v-w)! (m-a-b+u+v+w)! (2m)!}$ 
10 foreach row  $i$  in  $P$  do
11   if  $P(i) > 0$  then
12      $\phi(i) \leftarrow -P(i) \log_2 P(i)$ 
13   else
14      $\phi(i) \leftarrow 0$ 
15 Entropy  $h_{RSM} \leftarrow$  sum over all elements in  $\phi$ 
16 Max entropy  $Maxh_{RSM} \leftarrow \log_2$  of number of rows in  $UVW$ 
17  $Cov \leftarrow$  covariance matrix of  $UVW$ 
18 Entropy approximation  $Apxh_{RSM} \leftarrow \log_2 \left( \sqrt{2\pi e Cov} \right)$ 
19  $Q_{aa} \leftarrow \frac{a(a-1)}{2m(2m-1)}$ ,  $Q_{bb} \leftarrow \frac{b(b-1)}{2m(2m-1)}$ , and  $Q_{ab} \leftarrow \frac{2ab}{2m(2m-1)}$ 
20  $Q_c \leftarrow (1 - Q_{aa} - Q_{bb} - Q_{ab})$ 
21  $multprobs \leftarrow [Q_{aa} \ Q_{bb} \ Q_{ab} \ Q_c]$ 
22  $U \leftarrow 0$  to  $m$ ,  $V \leftarrow 0$  to  $m$ , and  $W \leftarrow 0$  to  $m$ 
23 Produce three-dimensional coordinate arrays where the output coordinate arrays  $u$ ,  $v$ , and  $w$  contain copies of the grid vectors  $U$ ,  $V$ , and  $W$ , respectively
24  $x \leftarrow m - (u + v + w)$ 
25  $\mathbf{X} \leftarrow [u \ v \ w \ x]$ 
26 Remove rows in  $\mathbf{X}$  that have negative elements and check that there are  $\binom{m+3}{3}$  rows left
27 for  $i \leftarrow 0$  to  $\binom{m+3}{3}$  do
28    $B(i+1) \leftarrow$  the pdf for the multinomial distribution with probabilities  $multprobs$  evaluated at each row
29    $\mathbf{X}(i+1)$ 
29 for  $i \leftarrow 0$  to  $\binom{m+3}{3}$  do
30   if  $B(i) > 0$  then
31      $\phi(i) \leftarrow -B(i) \log_2 B(i)$ 
32   else
33      $\phi(i) \leftarrow 0$ 
34 Entropy  $h_{IEA} \leftarrow$  sum over all elements in  $\phi$ 
35 Max entropy  $Maxh_{IEA} \leftarrow \log_2 \binom{m+3}{3}$ 
36 Entropy approximation  $Apxh_{IEA} \leftarrow \log_2 \left( \sqrt{(2\pi em)^3 Q_{aa} Q_{bb} Q_{ab} Q_c} \right)$ 
37 return  $h_{RSM}$ ,  $Maxh_{RSM}$ ,  $Apxh_{RSM}$ ,  $h_{IEA}$ ,  $Maxh_{IEA}$ ,  $Apxh_{IEA}$ 

```

Algorithm 6: Entropies and moments of marginal loop distributions under RSM and IEA

Input: Number of edges m
Output: The entropies h and entropy approximations $Apxh$ of $P(m_{ii} = v)$ under RSM and IEA, together with expected values and variances, respectively

```

1 for  $i \leftarrow 2$  to  $2m$  do
2    $\lfloor$   $deg(i) = i$ 
3 foreach row  $i$  in  $deg$  do
4    $d = deg(i)$ 
5   for  $v \leftarrow 0$  to  $\lfloor d/2 \rfloor$  do
6     if  $(m - d + v) < 0$  then
7        $\lfloor$   $P(i, v + 1) \leftarrow 0$ 
8     else
9        $\lfloor$   $P(i, v + 1) \leftarrow (m! 2^{d-2v} d! (2m - d)! / (v! (d - 2v)! (m - d + v)! 2m!)$ 
10     $Q \leftarrow d(d - 1) / (2m(2m - 1))$ 
11    for  $v \leftarrow 0$  to  $m$  do
12       $\lfloor$   $B(i, v + 1) \leftarrow$  the pdf for the binomial distribution with parameters  $Q$  and  $m$  evaluated at point  $v$ 
13     $\mu(i) \leftarrow d(d - 1) / (2(2m - 1))$ 
14     $\sigma^2(i) \leftarrow \mu(i)(1 - \mu(i)/m)$ 
15     $\Delta(i) \leftarrow (d(d - 1)(d - 2)(d - 3)) / (4(2m - 1)(2m - 3)) - (\mu(i)^2((m - 1)/m))$ 
16     $Var \leftarrow \sigma^2(i) + \Delta(i)$ 
17 foreach row  $i$  in  $P$  do
18   foreach column  $j$  in  $P$  do
19     if  $P(i, j) > 0$  then
20        $\lfloor$   $\phi(i, j) \leftarrow -P(i, j) \log_2 P(i, j)$ 
21     else
22        $\lfloor$   $\phi(i, j) \leftarrow 0$ 
23   Entropy  $h_{RSM}(i) \leftarrow$  sum over all columns in  $\phi(i)$ 
24 foreach row  $i$  in  $B$  do
25   foreach column  $j$  in  $B$  do
26     if  $B(i, j) > 0$  then
27        $\lfloor$   $\phi(i, j) \leftarrow -B(i, j) \log_2 B(i, j)$ 
28     else
29        $\lfloor$   $\phi(i, j) \leftarrow 0$ 
30   Entropy  $h_{IEA}(i) \leftarrow$  sum over all columns in  $\phi(i)$ 
31 foreach row  $i$  in  $deg$  do
32   Entropy approximation RSM  $Apxh_{RSM}(i) \leftarrow \log_2 \left( \sqrt{2\pi e Var(i)} \right)$ 
33   Entropy approximation IEA  $Apxh_{IEA}(i) \leftarrow \log_2 \left( \sqrt{2\pi e \sigma^2(i)} \right)$ 
34 return List with columns containing  $deg, h_{RSM}, h_{IEA}, Apxh_{RSM}, Apxh_{IEA}, \mu, \sigma^2, \Delta, Var$ 

```

Algorithm 7: Entropies and moments of marginal non-loop distributions under RSM and IEA

Input: Number of edges m
Output: The entropies h and entropy approximations $Apxh$ of $P(m_{ij} = w)$ under RSM and IEA, together with expected values and variances, respectively

```

1 for  $i \leftarrow 2$  to  $m$  do
2   for  $j \leftarrow i$  to  $2m - i$  do
3     Each row in  $deg \leftarrow [i \ j]$ 
4   foreach row  $i$  in  $deg$  do
5      $d_i \leftarrow deg(i, 1)$ 
6      $d_j \leftarrow deg(i, 2)$ 
7     Call Algorithm 5 for  $P$ 
8      $a \leftarrow d_i$ ,  $b \leftarrow d_j$  and  $c \leftarrow \min(a, b)$ 
9     for  $w \leftarrow 0$  to  $c$  do
10       $P_w(i, w + 1) \leftarrow$  sum over  $a$  to  $\lfloor (a - w)/2 \rfloor$  and  $b$  to  $\lfloor (b - w)/2 \rfloor$  in  $P$ 
11       $Q \leftarrow 2ab/(2m(2m - 1))$ 
12      for  $w \leftarrow 0$  to  $m$  do
13         $B_w(i, w + 1) \leftarrow$  the pdf for the binomial distribution with parameters  $Q$  and  $m$  evaluated at point  $w$ 
14         $\mu(i) \leftarrow 2ab/(2(2m - 1))$ 
15         $\sigma^2(i) \leftarrow \mu(i)(1 - \mu(i)/m)$ 
16         $\Delta(i) \leftarrow (ab(a - 1)(b - 1)) / ((2m - 1)(2m - 3)) - (\mu(i)^2((m - 1)/m))$ 
17         $Var \leftarrow \sigma^2(i) + \Delta(i)$ 
18      foreach row  $i$  in  $P_w$  do
19        foreach column  $j$  in  $P_w$  do
20          if  $P_w(i, j) > 0$  then
21             $\phi(i, j) \leftarrow -P_w(i, j) \log_2 P_w(i, j)$ 
22          else
23             $\phi(i, j) \leftarrow 0$ 
24        Entropy  $h_{RSM}(i) \leftarrow$  sum over all columns in  $\phi(i)$ 
25      foreach row  $i$  in  $B_w$  do
26        foreach column  $j$  in  $B_w$  do
27          if  $B_w(i, j) > 0$  then
28             $\phi(i, j) \leftarrow -B_w(i, j) \log_2 B_w(i, j)$ 
29          else
30             $\phi(i, j) \leftarrow 0$ 
31        Entropy  $h_{IEA}(i) \leftarrow$  sum over all columns in  $\phi(i)$ 
32      foreach row  $i$  in  $D$  do
33        Entropy approximation RSM  $Apxh_{RSM}(i) \leftarrow \log_2 \left( \sqrt{2\pi e Var(i)} \right)$ 
34        Entropy approximation IEA  $Apxh_{IEA}(i) \leftarrow \log_2 \left( \sqrt{2\pi e \sigma^2(i)} \right)$ 
35 return List with columns containing  $deg$ ,  $h_{RSM}$ ,  $h_{IEA}$ ,  $Apxh_{RSM}$ ,  $Apxh_{IEA}$ ,  $\mu$ ,  $\sigma^2$ ,  $\Delta$ ,  $Var$ 

```

Algorithm 8: Information divergence between trivariate edge multiplicity distributions under RSM and IEA

Input: Number of edges m , degrees d_i and d_j at vertices i and j
Output: The information divergence D between $P((m_{ii}, m_{jj}, m_{ij})) = (u, v, w)$ under RSM and IEA

- 1 Call **Algorithm 5** for UVW , P and $multprobs$
- 2 **foreach** row i in UVW **do**
- 3 $UVW(i, 4) \leftarrow (m - UVW(i, 1) - UVW(i, 2) - UVW(i, 3))$
- 4 **foreach** row i in UVW **do**
- 5 $B(i) \leftarrow$ the pdf for the multinomial distribution with probabilities $multprobs$ evaluated at each row
 $UVW(i)$
- 6 **foreach** row i in UVW **do**
- 7 Weighted log-likelihood ratio $LLR(i) \leftarrow P_{uvw}(i) \log_2(P(i)/B(i))$
- 8 Divergence $D \leftarrow$ sum over all elements in LLR
- 9 **return** D

Algorithm 9: Information divergence between marginal loop multiplicity distributions under RSM and IEA

Input: Number of edges m
Output: The information divergence D between $P(m_{ii} = v)$ under RSM and IEA

- 1 Call **Algorithm 6** for deg , P and B
- 2 **foreach** row i in P **do**
- 3 **foreach** column j in P **do**
- 4 **if** $P(i, j) > 0$ **then**
- 5 Weighted log-likelihood ratio $LLR(j) \leftarrow P(i, j) \log_2(P(i, j)/B(i, j))$
- 6 **else**
- 7 $LLR(j) \leftarrow 0$
- 8 Divergence $D(i) \leftarrow$ sum over all elements $LLR(j)$
- 9 **return** List with columns containing deg , D

Algorithm 10: Information divergence between marginal non-loop multiplicity distributions under RSM and IEA

Input: Number of edges m
Output: The information divergence D between $P(m_{ij} = w)$ under RSM and IEA

- 1 Call **Algorithm 7** for deg , P_w and B_w
- 2 **foreach** row i in P_w **do**
- 3 **foreach** column j in P_w **do**
- 4 **if** $P_w(i, j) > 0$ **then**
- 5 Weighted log-likelihood ratio $LLR(j) \leftarrow P_w(i, j) \log_2(P_w(i, j)/B_w(i, j))$
- 6 **else**
- 7 $LLR(j) \leftarrow 0$
- 8 Divergence $D(i) \leftarrow$ sum over all elements $LLR(j)$
- 9 **return** List with columns containing deg , D

Algorithm 11: Entropies of and information divergence between distributions of multigraphs under RSM and IEA

Input: Degree or stub multiplicity sequence $\mathbf{d} = (d_1, d_2, \dots, d_n)$
Output: The entropies h , entropy upper bounds $Maxh$, entropy approximation $Apxh$ and information divergence D of and between the multigraph distributions under RSM and IEA

- 1 Number of vertices $n \leftarrow$ number of columns in \mathbf{d}
- 2 Number of edges $m \leftarrow$ half of the sum of all element values in \mathbf{d}
- 3 Call **Algorithm 4** for \mathbf{m} , $P_{\mathbf{Z}}$ and $K_{\mathbf{d}}$
- 4 **foreach** row i in $P_{\mathbf{Z}}$ **do**
- 5 **if** $P_{\mathbf{Z}}(i) > 0$ **then**
- 6 $\phi(i) \leftarrow -P_{\mathbf{Z}}(i) \log_2 P_{\mathbf{Z}}(i)$
- 7 **else**
- 8 $\phi(i) \leftarrow 0$
- 9 Entropy $h_{RSM} \leftarrow$ sum over all elements in ϕ
- 10 Max entropy $Maxh_{RSM} \leftarrow \log_2 K_{\mathbf{d}}$
- 11 $CovRSM \leftarrow$ covariance matrix of \mathbf{m}
- 12 Entropy approximation $Apxh_{RSM} \leftarrow \log_2 \left(\sqrt{\det(2\pi e CovRSM)} \right)$
- 13 **for** $i \leftarrow 0$ **to** n **do**
- 14 **for** $j \leftarrow 0$ **to** n **do**
- 15 **if** $i = j$ **then**
- 16 $\mathbf{Q}(i, j) \leftarrow (\mathbf{d}(i)(\mathbf{d}(i) - 1)) / (2m(2m - 1))$
- 17 **else if** $i < j$ **then**
- 18 $\mathbf{Q}(i, j) \leftarrow 2\mathbf{d}(i)\mathbf{d}(j) / (2m(2m - 1))$
- 19 **else**
- 20 $\mathbf{Q}(i, j) \leftarrow 0$
- 21 $Q \leftarrow$ vector containing the upper triangular elements $i \leq j$ of $\mathbf{Q}(i, j)$
- 22 $r \leftarrow \binom{n+1}{2}$
- 23 **for** $i \leftarrow 1$ **to** $r - 1$ **do**
- 24 **for** $j \leftarrow 1$ **to** $r - 1$ **do**
- 25 **if** $i = j$ **then**
- 26 $CovIEA(i, j) \leftarrow mQ(i)(1 - Q(i))$
- 27 **else**
- 28 $CovIEA(i, j) \leftarrow -mQ(i)Q(j)$
- 29 Max entropy $Maxh_{IEA} \leftarrow \log_2 \binom{m+r-1}{m}$
- 30 Entropy approximation $Apxh_{IEA} \leftarrow \log_2 \left(\sqrt{\det(2\pi e CovIEA)} \right)$
- 31 **for** $i \leftarrow 1$ **to** m **do**
- 32 $P_{IEA}(i) \leftarrow$ the pdf for the multinomial distribution with probabilities Q evaluated at each row
 $\mathbf{m}(i)$
- 33 **foreach** row i in $P_{\mathbf{Z}}$ **do**
- 34 Weighted log-likelihood ratio $LLR(i) \leftarrow P_{\mathbf{Z}}(i) \log_2 (P_{\mathbf{Z}}(i) / P_{IEA}(i))$
- 35 Divergence $D \leftarrow$ sum over all elements in LLR
- 36 **return** $h_{RSM}, Maxh_{RSM}, Apxh_{RSM}, Maxh_{IEA}, Apxh_{IEA}, D$

Algorithm 12: Entropy approximations of distributions of multigraphs under RSM

Input: Degree or stub multiplicity sequence $\mathbf{d} = (d_1, d_2, \dots, d_n)$

Output: Entropy H under RSM, and the two entropy approximations H^* and H^{**} based on expected entropy and asymptotic entropy under ISA

```

1 Number of vertices  $n \leftarrow$  number of columns in  $\mathbf{d}$ 
2 Number of edges  $m \leftarrow$  half of the sum of all element values in  $\mathbf{d}$ 
3 Call Algorithm 11 for  $h_{RSM}$ 
4  $H \leftarrow h_{RSM}$ 
5  $prod \leftarrow$  product over all the elements in  $\mathbf{d}$ 
6 if  $n > 2$  then
7    $H^* \leftarrow \log_2 \left( \sqrt{(2\pi em)^{\binom{n}{2}} 2^{\binom{n-1}{2}} prod^n (1/2m)^{n^2}} \right)$ 
8 else
9    $H^* \leftarrow \log_2 \left( \sqrt{(2\pi em)^{\binom{n}{2}} prod^n (1/2m)^{n^2}} \right)$ 
10  $\mathbf{p} \leftarrow$  vector containing each element in  $\mathbf{d}$  divided by  $2m$ 
11 for  $i \leftarrow 1$  to  $n$  do
12    $x(i) \leftarrow -\mathbf{p}(i) \log_2 \mathbf{p}(i)$ 
13  $h_{\mathbf{p}} \leftarrow$  sum over all elements in  $x$ 
14  $n_c \leftarrow 2^{h_{\mathbf{p}}}$ 
15  $r_c \leftarrow n_c^2 2^{-(n_c-1)/n_c}$ 
16  $a^{**} \leftarrow \log_2 \left( \sqrt{(2\pi e)^{r_c-1} (n_c^{n_c}) / (4\pi e)^{n_c-1} r_c^{r_c}} \right)$ 
17  $b^{**} \leftarrow (r_c - n_c) / 2$ 
18  $H^{**} \leftarrow a^{**} + b^{**} \log_2(m)$ 
19 return  $H, H^*, H^{**}$ 

```

Algorithm 13: Approximations of the probability that an RSM multigraph is simple

Input: Degree or stub multiplicity sequence $\mathbf{d} = (d_1, d_2, \dots, d_n)$
Output: The probability that an RSM multigraph is simple P_{RSM} , together with the three approximations P_{1Asymp} , P_{2Asymp} and $P_{Poisson}$

- 1 Number of vertices $n \leftarrow$ number of columns in \mathbf{d}
- 2 Number of edges $m \leftarrow$ half of the sum of all element values in \mathbf{d}
- 3 Call **Algorithm 4** for \mathbf{m} , m_1 , I_{m_2} and $K_{\mathbf{d}}$
- 4 **foreach** row i in \mathbf{m} **do**
- 5 **if** $m_1(i) = 0$ **and** $I_{m_2}(i) = 0$ **then**
- 6 $simple(i) \leftarrow 1$
- 7 **else**
- 8 $simple(i) \leftarrow 0$
- 9 $P_{RSM} \leftarrow$ sum over all elements in $simple$ divided by $K_{\mathbf{d}}$
- 10 $S \leftarrow$ sum over all squared elements in \mathbf{d}
- 11 $P_{1Asymp} \leftarrow \exp(-1/4(S/2m)^2 + 1/4)$
- 12 **for** $i \leftarrow 0$ **to** n **do**
- 13 **for** $j \leftarrow 0$ **to** n **do**
- 14 **if** $i = j$ **then**
- 15 $L(i, j) \leftarrow \mathbf{d}(i)(\mathbf{d}(i) - 1)/2m$
- 16 **else if** $i < j$ **then**
- 17 $L(i, j) \leftarrow \sqrt{\mathbf{d}(i)(\mathbf{d}(i) - 1)\mathbf{d}(j)(\mathbf{d}(j) - 1)}/2m$
- 18 **else**
- 19 $L(i, j) \leftarrow 0$
- 20 $X \leftarrow L - \log(1 + L)$
- 21 $Y \leftarrow$ matrix with elements above the diagonal in X
- 22 $sum1 \leftarrow$ sum over the diagonal in L
- 23 $sum2 \leftarrow$ sum over all rows and columns in Y
- 24 $P_{2Asymp} \leftarrow \exp(-1/2(sum1) - sum2)$
- 25 **for** $i \leftarrow 0$ **to** n **do**
- 26 **for** $j \leftarrow 0$ **to** n **do**
- 27 **if** $i = j$ **then**
- 28 $\mathbf{Q}(i, j) \leftarrow (\mathbf{d}(i)(\mathbf{d}(i) - 1))/(2m(2m - 1))$
- 29 **else if** $i < j$ **then**
- 30 $\mathbf{Q}(i, j) \leftarrow 2\mathbf{d}(i)\mathbf{d}(j)/(2m(2m - 1))$
- 31 **else**
- 32 $\mathbf{Q}(i, j) \leftarrow 0$
- 33 **for** $i \leftarrow 0$ **to** n **do**
- 34 **for** $j \leftarrow 0$ **to** n **do**
- 35 **if** $i < j$ **then**
- 36 $\mathbf{bin}(i, j) \leftarrow$ the pdf for the binomial distribution with parameters $Q(i, j)$ and m evaluated at point 1
- 37 $\lambda \leftarrow$ sum over all columns and rows in \mathbf{bin}
- 38 $P_{Poisson} \leftarrow$ the pdf of the Poisson distribution with parameter λ evaluated at point m
- 39 **return** P_{RSM} , P_{1Asymp} , P_{2Asymp} , $P_{Poisson}$

Algorithm 14: Distribution of the multiplicity sequence under an RSM, IEAS or ISA model

Input: Model $RSM(\mathbf{d})$, $IEAS(\mathbf{d})$ or $ISA(\mathbf{p})$, where \mathbf{d} and \mathbf{p} are specified
Output: Multiplicity sequences \mathbf{m} and their probabilities $probs$ under the specified model

```

1  $r \leftarrow \binom{n+1}{2}$ 
2 if  $Model=RSM(\mathbf{d})$  then
3   Number of vertices  $n \leftarrow$  number of columns in  $\mathbf{d}$ 
4   Number of edges  $m \leftarrow$  half of the sum of all element values in  $\mathbf{d}$ 
5   Call Algorithm 4 for  $\mathbf{m}$  and  $P_{\mathbf{Z}}$ 
6    $probs \leftarrow P_{\mathbf{Z}}$ 
7 else if  $Model=IEAS(\mathbf{d})$  then
8   Number of vertices  $n \leftarrow$  number of columns in  $\mathbf{d}$ 
9   Number of edges  $m \leftarrow$  half of the sum of all element values in  $\mathbf{d}$ 
10  Call Algorithm 11 for  $\mathbf{Q}$ 
11   $Q \leftarrow$  vector containing the upper triangular elements  $i \leq j$  of  $\mathbf{Q}(i, j)$ 
12   $\mathbf{m} \leftarrow \binom{m+r-1}{m}$  rows of possible multiplicity sequences under IEAS
13  foreach row  $i$  in  $\mathbf{m}$  do
14     $probs(i) \leftarrow$  the pdf for the multinomial distribution with probabilities  $Q$  evaluated at each
    row  $\mathbf{m}(i)$ 
15 else if  $Model=ISA(\mathbf{p})$  then
16   Number of vertices  $n \leftarrow$  number of columns in  $\mathbf{p}$ 
17   Number of edges  $m \leftarrow$  half of the sum of all element values in  $\mathbf{p}$ 
18   for  $i \leftarrow 0$  to  $n$  do
19     for  $j \leftarrow 0$  to  $n$  do
20       if  $i = j$  then
21          $\mathbf{Q}(i, j) \leftarrow \mathbf{p}(i)^2$ 
22       else if  $i < j$  then
23          $\mathbf{Q}(i, j) \leftarrow 2\mathbf{p}(i)\mathbf{p}(j)$ 
24       else
25          $\mathbf{Q}(i, j) \leftarrow 0$ 
26    $Q \leftarrow$  vector containing the upper triangular elements  $i \leq j$  of  $\mathbf{Q}(i, j)$ 
27    $\mathbf{m} \leftarrow \binom{m+r-1}{m}$  rows of possible multiplicity sequences under ISA
28   foreach row  $i$  in  $\mathbf{m}$  do
29      $probs(i) \leftarrow$  the pdf for the multinomial distribution with probabilities  $Q$  evaluated at each
    row  $\mathbf{m}(i)$ 
30 return  $\mathbf{m}$ ,  $probs$ 

```

Algorithm 15: Statistical tests of a simple IEA hypothesis

Input: Model $RSM(\mathbf{d})$, $IEAS(\mathbf{d})$ or $ISA(\mathbf{p})$, and hypothesis $IEAS(\mathbf{d}_0)$ or $ISA(\mathbf{p}_0)$ where \mathbf{d} , \mathbf{d}_0 , \mathbf{p} and \mathbf{p}_0 are specified

Output: Outcomes of the Pearson goodness-of-fit statistic S with probabilities P_S , and outcomes of the divergence statistic T with probabilities P_T

```

1 Call Algorithm 14 for  $\mathbf{m}$  and probs according to input model
2 if Hypothesis= $IEAS(\mathbf{d}_0)$  then
3    $\mathbf{d} \leftarrow \mathbf{d}_0$ 
4   Call Algorithm 11 for  $\mathbf{Q}$ 
5    $Q \leftarrow$  vector containing the upper triangular elements  $i \leq j$  of  $\mathbf{Q}(i, j)$ 
6    $\mathbf{m} \leftarrow \binom{m+r-1}{m}$  rows of possible multiplicity sequences under IEAS
7   foreach row  $i$  in  $\mathbf{m}$  do
8      $O_S(i) \leftarrow \mathbf{m}(i)$ ,  $E_S(i) \leftarrow mQ$ 
9     foreach column  $j$  in  $\mathbf{m}$  do
10      if  $E_S(j) = 0$  then
11         $x(j) \leftarrow 0$ 
12      else
13         $x(j) \leftarrow (O_S(j) - E_S(j))^2 / E_S(j)$ 
14       $S(i) \leftarrow$  sum over all columns in  $x$ 
15   $S_{UNI} \leftarrow$  unique values of  $S$ 
16  foreach row  $i$  in  $S$  do
17     $P_S(i) \leftarrow$  sum over all probs where  $S_{UNI} = S(i)$ 
18  foreach row  $i$  in  $\mathbf{m}$  do
19     $O_D(i) \leftarrow \mathbf{m}(i)$ ,  $E_D(i) \leftarrow mQ$ 
20    foreach column  $j$  in  $\mathbf{m}(i)$  do
21      if  $O_D(j) > 0$  and  $E_D(j) > 0$  then
22         $x(j) \leftarrow (O_D(j)/m) \log_2(O_D(j)/E_D(j))$ 
23      else
24         $x(j) \leftarrow 0$ 
25     $D(i) \leftarrow$  sum over all columns in  $x$ 
26     $T(i) \leftarrow 2mD(i) / \log_2(e)$ 
27   $T_{UNI} \leftarrow$  unique values of  $T$ 
28  foreach row  $i$  in  $T$  do
29     $P_T(i) \leftarrow$  sum over all probs where  $T_{UNI} = T(i)$ 
30 else if Hypothesis= $ISA(\mathbf{p}_0)$  then
31   for  $i \leftarrow 0$  to  $n$  do
32     for  $j \leftarrow 0$  to  $n$  do
33       if  $i = j$  then
34          $\mathbf{Q}(i, j) \leftarrow \mathbf{p}_0(i)^2$ 
35       else if  $i < j$  then
36          $\mathbf{Q}(i, j) \leftarrow 2\mathbf{p}_0(i)\mathbf{p}_0(j)$ 
37       else
38          $\mathbf{Q}(i, j) \leftarrow 0$ 
39   Repeat steps 5-29
40 return  $S_{UNI}$ ,  $P_S$ ,  $T_{UNI}$ ,  $P_T$ 

```

Algorithm 16: Statistical tests of a composite IEAS hypothesis

Input: Model $RSM(\mathbf{d})$ IEAS(\mathbf{d}) or ISA(\mathbf{p}) and hypothesis IEAS

Output: Outcomes of the Pearson goodness-of-fit statistic S with probabilities P_S , and outcomes of the divergence statistic T with probabilities P_T

```
1 Call Algorithm 14 for  $\mathbf{m}$  and probs according to input model
2 foreach row  $i$  in  $\mathbf{m}$  do
3    $M \leftarrow$  upper triangular matrix containing the elements in  $\mathbf{m}(i)$ 
4    $M \leftarrow M + M'$ 
5    $\mathbf{d}_{EST}(i) \leftarrow$  sum over all rows (or columns) in  $M$ 
6    $\mathbf{d} \leftarrow \mathbf{d}_{EST}(i)$ 
7   Call Algorithm 11 for  $\mathbf{Q}$ 
8    $Q \leftarrow$  vector containing the upper triangular elements  $i \leq j$  of  $\mathbf{Q}(i, j)$ 
9    $O_S(i) \leftarrow \mathbf{m}(i)$ 
10   $E_S(i) \leftarrow mQ$ 
11  foreach column  $j$  in  $\mathbf{m}$  do
12    if  $E_S(j) = 0$  then
13       $x(j) \leftarrow 0$ 
14    else
15       $x(j) \leftarrow (O_S(j) - E_S(j))^2 / E_S(j)$ 
16   $S(i) \leftarrow$  sum over all columns in  $x$ 
17   $O_D(i) \leftarrow \mathbf{m}(i)$ 
18   $E_D(i) \leftarrow mQ$ 
19  foreach column  $j$  in  $\mathbf{m}(i)$  do
20    if  $O_D(j) > 0$  and  $E_D(j) > 0$  then
21       $x(j) \leftarrow (O_D(j)/m) \log_2(O_D(j)/E_D(j))$ 
22    else
23       $x(j) \leftarrow 0$ 
24   $D(i) \leftarrow$  sum over all columns in  $x$ 
25   $T(i) \leftarrow 2mD(i) / \log_2(e)$ 
26  Clear  $M, \mathbf{d}, \mathbf{Q}, Q$ 
27  $S_{UNI} \leftarrow$  unique values of  $S$ 
28 foreach row  $i$  in  $S$  do
29    $P_S(i) \leftarrow$  sum over all probs where  $S_{UNI} = S(i)$ 
30  $T_{UNI} \leftarrow$  unique values of  $T$ 
31 foreach row  $i$  in  $T$  do
32    $P_T(i) \leftarrow$  sum over all probs where  $T_{UNI} = T(i)$ 
33 return  $S_{UNI}, P_S, T_{UNI}, P_T$ 
```

Algorithm 17: Statistical tests of a composite ISA hypothesis

Input: Model $RSM(\mathbf{d})$ $IEAS(\mathbf{d})$ or $ISA(\mathbf{p})$ and hypothesis ISA

Output: Outcomes of the Pearson goodness-of-fit statistic S with probabilities P_S , and outcomes of the divergence statistic T with probabilities P_T

```
1 Call Algorithm 14 for  $\mathbf{m}$  and  $probs$  according to input model
2 foreach row  $i$  in  $\mathbf{m}$  do
3    $M \leftarrow$  upper triangular matrix containing the elements in  $\mathbf{m}(i)$ 
4    $M \leftarrow M + M'$ 
5    $\mathbf{d}_{EST}(i) \leftarrow$  sum over all rows (or columns) in  $M$ 
6    $\mathbf{p}_{EST}(i) \leftarrow$  each element in  $\mathbf{d}_{EST}(i)$  divided by  $2m$ 
7   for  $i \leftarrow 0$  to  $n$  do
8     for  $j \leftarrow 0$  to  $n$  do
9       if  $i = j$  then
10         $\mathbf{Q}(i, j) \leftarrow \mathbf{p}_{EST}(i)^2$ 
11      else if  $i < j$  then
12         $\mathbf{Q}(i, j) \leftarrow 2\mathbf{p}_{EST}(i)\mathbf{p}_{EST}(j)$ 
13      else
14         $\mathbf{Q}(i, j) \leftarrow 0$ 
15   $Q(i) \leftarrow$  the upper triangular elements  $i \leq j$  of  $\mathbf{Q}(i, j)$ 
16  Clear  $M, \mathbf{Q}$ 
17 foreach row  $i$  in  $\mathbf{m}$  do
18    $Q \leftarrow Q(i)$ 
19   Repeat steps 9-25 in Algorithm 16
20 Repeat steps 27-32 in Algorithm 16
21 return  $S_{UNI}, P_S, T_{UNI}, P_T$ 
```

References

Frank, O. and Shafie, T. (2012), Complexity of Families of Multigraphs, to appear in *JSM Proceedings*, Section on Statistical Graphics, Alexandria, VA: American Statistical Association.

Shafie, T. (2012a), Random Stub Matching Models of Multigraphs, *Research Report 2012:1*, Department of Statistics, Stockholm University.

Shafie, T. (2012b), Statistical Analysis of Multigraphs, *Research Report 2012:2*, Department of Statistics, Stockholm University.