Random Multigraphs

Complexity Measures, Probability Models and Statistical Inference

Termeh Shafie

Contents

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. Various models have been proposed to study random graphs with fixed or modeled degrees (the number of edges incident to a vertex), degree distributions or expected degrees. The classical random graph introduced by Erdös and Rényi (1959, 1960) has independent edges and is fully symmetric with a common binomial distribution for the degree at any vertex. The Erdös-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 Erdös-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 independently, 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. Similarily, 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, $\mathbf{P} = (P_1, \ldots, P_N)$ and $\mathbf{Q} = (Q_1, \ldots, Q_N)$. 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.

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.

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.

Erdös, P. and Renyi, A. (1959), On Random Graphs, *Publicationes Mathematicae*, Debrecen, **6**, 290–297.

Erdös, 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.