Estimating Subgraph Generation Models

1st Laurens Bogaardt
Netherlands eScience Center
Amsterdam, the Netherlands
l.bogaardt@esciencecenter.nl

2nd Frank Takes
University of Amsterdam
Amsterdam, the Netherlands
takes@uva.nl

***Abstract*—Recently, a new network formation model was proposed. The current research looks into a method to estimate the parameters of this model based on the subgraph census.**

***Index Terms*—Networks, Graphs, ERGM, SUGM, Subgraphs**

I. Introduction

The Exponential Random Graph Model (ERGM) is the most frequently used network formation model. However, it suffers from two fundamental flaws [1]. Firstly, its parameter estimates are inconsistent. Secondly, it does not scale well. Recently, an alternative network formation model was suggested: the Subgraph Generation Model (SUGM) [1].

The original article describing SUGM contains two methods to estimate the parameters of the model. The current research suggests a third, more intuitive method based on the subgraph census. In an k-subgraph census, a network of n nodes is partitioned into all possible subsets of k nodes, which are then tallied according to their isomorphism class [2].

II. Subgraph Generation Model

A SUGM is defined by a set of l small subgraphs, such as links, triangles or stars, each with corresponding probabilities. For each subgraph i of m_i nodes, the n nodes of the entire network are partitioned into all possible subsets of m_i nodes. Then, each of these subsets receives the subgraph i with propability $1 - p_i$ or remains empty with probability p_i.

The observed network (left in Fig. 1) is the union of all these subgraphs (right in Fig. 1), where the generated subgraphs may overlap. Multiple neighbouring subgraphs may incidentally form additional structures such as triangles or squares.

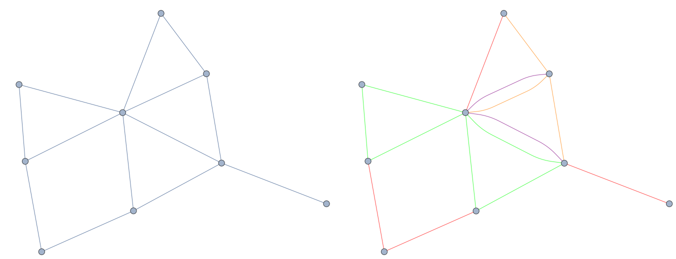


Fig. 1. The observed network (left) is the union (right) of randomly generated links (red), 2-paths (purple), triangles (green) and 3-stars (yellow).

Table I contains the probabilities of observing any of the possible triads for three different generation models. These probabilities enter into the multinomial probability mass function of (1), together with the counts of the census, to form the likelihood function. This can be used to estimate the parameters of the model and their confidence intervals.

$$f(x_1, \cdots, x_l; p_1, \cdots, p_l) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^{l} p_i^{x_i} \quad (1)$$

III. Further Research

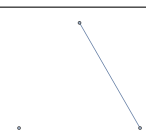
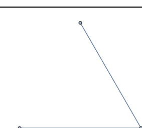
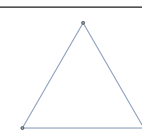
Future work should extend the list of possible subgraphs, deal with the correlations within the census, develop an *R*-package and apply the model to real-world data.

References

[1] A. G. Chandrasekhar and M. O. Jackson, "Tractable and consistent random graph models," *ArXiv*, 2014. [Online]. Available: https://arxiv.org/abs/1210.7375

[2] J. A. Davis and S. Leinhardt, "The structure of positive interpersonal relations in small groups," in *Sociological Theory in Progress*, J. Berger, M. Zelditch, and B. Anderson, Eds. Houghton-Mifflin, 1972.

TABLE I
Probabilities in the Subgraph Census

	Subgraphs of the Undirected Triad Census			
Model				
Links	p_L^3	$3\,p_L^2\,(1-p_L)$	$3\,p_L\,(1-p_L)^2$	$(1-p_L)^3$
Triangles	$p_T\,(p_T^{n-3})^3$	$3\,p_T\,(p_T^{n-3})^2\,(1-p_T^{n-3})$	$3\,p_T\,(p_T^{n-3})\,(1-p_T^{n-3})^2$	$(1-p_T)+p_T\,(1-p_T^{n-3})^3$
Links & Triangles	$p_T\,(p_L\,p_T^{n-3})^3$	$3\,p_T\,(p_L\,p_T^{n-3})^2\,(1-p_L\,p_T^{n-3})$	$3\,p_T\,(p_L\,p_T^{n-3})\,(1-p_L\,p_T^{n-3})^2$	$(1-p_T)+p_T\,(1-p_L\,p_T^{n-3})^3$