Estimating Subgraph Generation Models to Understand Large Network Formation

Laurens Bogaardt — Netherlands eScience Center

Frank W. Takes — University of Amsterdam

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.

Introduction

Social scientists often aim to understand the incentives and mechanisms which result in large scale socio-economic structures. Key to this is network formation analysis. However, large datasets are not uncommon, leading to a computational challenge. For example, political scientists interested in global networks of corporate control may need to analyse millions of companies to answer their questions [1, 2].

ERGM

The Exponential Random Graph Model (ERGM) is a frequently used network formation model. Unfortunately, it suffers from two fundamental flaws. Firstly, its parameter estimates are inconsistent [3, 4]. Secondly, it does not scale well [5]. Recently, an alternative network formation model was suggested: the Subgraph Generation Model (SUGM) [6, 7, 8].

SUGM

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 grouped into all possible subsets of m_i nodes. Then, each of these subsets receives the subgraph i with probability $1 - p_i$ or remains empty with probability p_i.
The observed network, depicted in Fig. 1, is the union of all these subgraphs, depicted in Fig. 2. The various generated subgraphs may have some overlapping edges. Multiple neighbouring subgraphs may incidentally form additional structures such as triangles or squares.

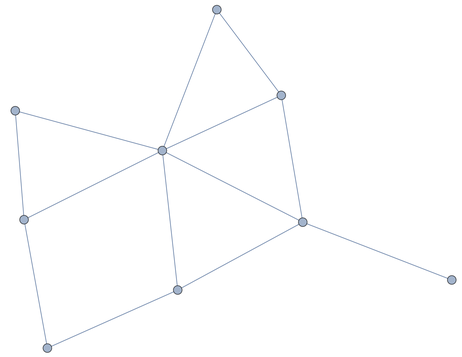

Figure 1: The observed network is the union of randomly generated subgraph.

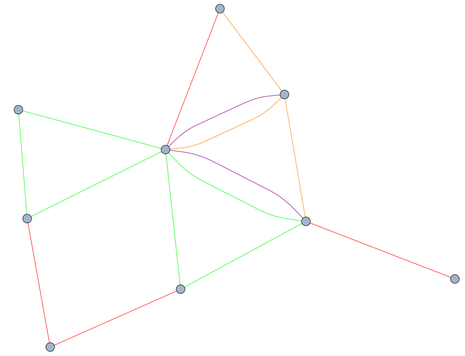

Figure 2: Randomly generated links (red), 2-paths (purple), triangles (green) and 3-stars (yellow).

Estimation

The original articles describing SUGM contain two methods to estimate the parameters of the model. The current research suggests a third, more intuitive method based on the subgraph census. In a k-subgraph census, a network of n nodes is grouped into all possible subsets of k nodes, which are then tallied according to their isomorphism class [9, 10, 11].
The table below contains the probabilities of observing any of the possible triads for three different generation models. In general, each of the r counts of the census x_j, together with the probability functions $f_j(\hat{p}_1, \cdots, \hat{p}_l)$, enter into the multinomial ... probability mass function of (1) to form the likelihood function. This can be used to estimate the parameters of the model and their confidence intervals.

$$\mathcal{L}(f_1, \cdots, f_r | x_1, \cdots, x_r) \sim \prod_{j=1}^{r} f_j^{x_j} \qquad (1)$$

Conclusion

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.

Probabilities in the Subgraph 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	$p_T\,(p_L\,p_T^{n-3})^3$	$3\,p_T\,(p_L\,p_T^{n-3})^2\,(1-p_L\,p_T^{n-3})$	$3\,p_T\,(p_L\,p_T^{n-3})\,(1-p_L\,p_T^{n-3})^2$	$(1-p_T)+p_T\,(1-p_L\,p_T^{n-3})^3$

References

[1] Frank W. Takes and Eelke M. Heemskerk. Centrality in the global network of corporate control. *Social Network Analysis and Mining*, 6:1–18, 2016.

[2] Meindert Fennema and Eelke M. Heemskerk. When theory meets methods: the naissance of computer assisted corporate interlock research. *Global Networks*, 1:81–104, 2018.

[3] Cosma Rohilla Shalizi and Alessandro Rinaldo. Consistency under sampling of exponential random graph models. *The Annals of Statistics*, 41:508–535, 2013.

[4] Sourav Chatterjee and Persi Diaconis. Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41:2428–2461, 2013.

[5] Shankar Bhamidi, Guy Bresler, and Allan Sly. Mixing time of exponential random graphs. *The Annals of Applied Probability*, 21:2146–2170, 2011.

[6] Arun G. Chandrasekhar and Matthew O. Jackson. Tractable and consistent random graph models. *ArXiv*, 2014.

[7] Arun G. Chandrasekhar and Matthew O. Jackson. A network formation model based on subgraphs. *ArXiv*, 2016.

[8] Arun G. Chandrasekhar. Econometrics of network formation. In Yann Bramoulle, Andrea Galeotti, and Brian Rogers, editors, *The Oxford Handbook of the Economics of Networks*. Oxford University Press, 2016.

[9] James A. Davis and Samuel Leinhardt. The structure of positive interpersonal relations in small groups. In Joseph Berger, Morris Zelditch, and Bo Anderson, editors, *Sociological Theory in Progress*. Houghton-Mifflin, 1972.

[10] Paul W. Holland and Samuel Leinhardt. A method for detecting structure in sociometric data. *American Journal of Sociology*, 76:492–513, 1970.

[11] Paul W. Holland and Samuel Leinhardt. Local structure in social networks. *Sociological Methodology*, 7:1–45, 1976.

l.bogaardt@esciencecenter.nl | takes@uva.nl