Author: Li-Chun Zhang (University of Southampton)

Title: Sampling and design-based inference in finite networks

Co-authors: Melike Oguz-Alper

Abstract:

Many data that contain non-nested relationships and associated measures can be represented by a valued graph consisting of nodes and edges, i.e. network. Such data can naturally arise from real-world social, technological, biological or information networks. They can also result from combination of multiple data sources, such as when respective datasets of persons and businesses are joined together. The network data raises the question of what the best use of such data is in making statistical inference about the phenomenon of interest. There is an extensive literature on model-based analysis of networks. However, the modelling approach may not always be successful, because the underlying dynamics may be too complicated or transient or subject to shocks, etc. There is thus always scope for a design-based approach to finite networks.

In the conventional survey sampling theory, the population is envisaged as a list of units, possibly in a nested clustering structure, which is associated with a set of measures that are treated as fixed constants. The design-based inference refers to hypothetical repeated sampling from the given population under a well-defined probability design. There exist well-established techniques for finite-population sampling and estimation as such. Despite a few notable exceptions in the past, such as multiplicity sampling including indirect sampling, adaptive cluster sampling, the theory of sampling and inference for finite networks is relatively under-developed, and the methods are rarely applied in Official Statistics.

Zhang and Patone (2017) synthesise and extend the graph sampling theory, which covers all the existing network sampling techniques as special cases. They develop a general Horvitz-Thompson estimation approach under arbitrary T-stage snowball sampling. In this work we consider design-based inference for target parameters beyond the totals of measures associated the nodes. Examples of such higher-order network parameters are network density, reciprocity, transitivity, etc. We establish generally the relative efficiency of two types of Horvitz-Thompson estimators that exist in the literature of network sampling. An application to a labour flow network will be presented, where labour flows between industrial sectors form the edges, which are based on the Norwegian Income and Employment data in the administrative sources.