

STATISTICS COLLOQUIUM

MONDAY, APRIL 30, 2012

TALK: 4:00 PM — SCIENCE CENTER RM. 705

RECEPTION: 5:15 PM — SCIENCE CENTER, 7TH FLOOR

“Respondent-Driven Sampling and Homophily in Network Data”

Sergiy Nesterko
Statistics Department
Harvard University

ABSTRACT

Data that can be represented as a network, where there are measurements both on units and on pairs of units, are becoming increasingly prevalent in the social sciences and public health. Homophily in network data or the tendency of units to connect based on similar nodal attribute values (i.e. income, HIV status) more often than expected by chance is receiving strong attention from researchers in statistics, medicine, sociology, public health and others. Respondent-Driven Sampling (RDS) is a link-tracing network sampling strategy heavily used in public health worldwide that is cost efficient and allows us to survey populations inaccessible by conventional techniques.

Via extensive simulation we study the performance of existing methods of estimating population averages, and show that they have poor performance if there is homophily on the quantity surveyed. We propose the first model-based approach for this setting and show its superiority as a point estimator and in terms of uncertainty intervals coverage rates, and demonstrate its application to a real life RDS-based survey. We study how the strength of homophily effects can be estimated and compared across networks and different binary attributes under several network sampling schemes. We give a proof that homophily can be effectively estimated under RDS and propose a new homophily index. This work moves towards a deeper understanding of network structure as a function of nodal attributes and network sampling under homophily.

This is joint work with my advisor Joe Blitzstein.