Nested Case Control Sampling for Egocentric and Relational Cox Models

Duy Vu and David Hunter

Department of Statistics, The Pennsylvania State University

This work is supported by ONR under the MURI program, Award Number N00014-08-1-1015

1. Motivation

- Partial likelihoods for egocentric and relational Cox models:

$$PL(\beta) = \prod_{e=1}^{E} \left\{ \frac{exp[\beta' s(i_e, t_e)]}{\sum_{j \in R(t_e)} exp[\beta' s(j, t_e)]} \right\}$$

  - E is the number of events during the observation time.
  - $R(t_e)$ is the set of nodes or edges at risk at time t_e.
  - $s(j, t_e)$ is the covariate vector of node or edge j.
  - Running times of parameter β estimation algorithms are $O(EN)$ and $O(EN^2)$, respectively where N is the number of nodes.

- Factors N and N^2 can be reduced significantly if network covariates under consideration allow for sparse updates of sum denominators.

- For larger networks and richer sets of covariates, we need to consider other approaches including sampling-based and online inference methods.

2. The Risk Set Sampling Framework

- We denote the network history up to, but not including, time t as H_{t^-}.

- At time t_e when an event occurs on node or edge i_e (case), based on information in H_{t^-} we will sample a subset $\tilde{R}(t_e)$ (controls) from the current risk set $R(t_e)$. The case is always included in the sampled risk set.

- The modified partial likelihoods [Borgan et al, 1995]:

$$PL_s(\beta) = \prod_{e=1}^{E} \left\{ \frac{exp[\beta' s(i_e, t_e)] w_{i_e}(t_e)}{\sum_{j \in \tilde{R}(t_e)} exp[\beta' s(j, t_e)] w_j(t_e)} \right\}$$

  - Each sampled individual is weighted by $w_j(t_e)$ to compensate for differences in sampling probabilities.
  - Variant sampling designs based on H_{t^-} will results in different definitions of $w_j(t_e)$.
  - Packages for weighted conditional logistic regression models can be used for parameters estimation.

- In the $1{:}m$ nested case-control sampling design, each sampled risk set will contain the case and $m-1$ controls which are randomly sampled (without replacement) from the current risk set, i.e. $w_j(t_e)$ are equal for all j.

3. Egocentric Twitter-Vaccine Data

- We consider a network of 101,853 Twitter users from 12/5/2009 to 1/19/2010 [Salathé et al, 2011].

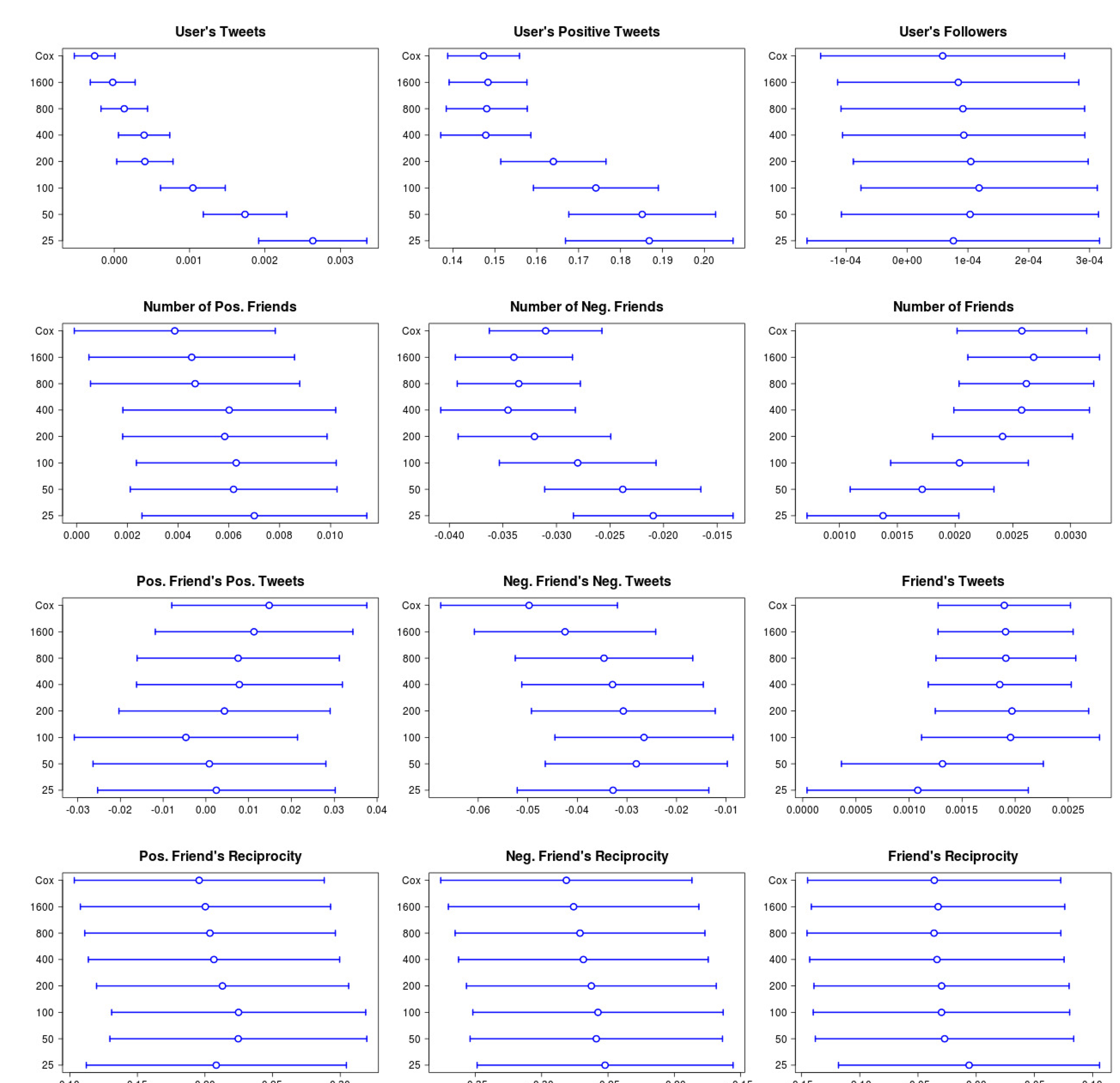
- There are 4,619,852 following edges among these users who have made 53,300 tweets about influenza A(H1N1) vaccine.

- Each tweet is classified as + (6,416), - (3,510), or neutral (43,374).

- We are interested in how the *positive* sentiment of future tweets of a user is associated with her past tweeting behavior as well as her friends' ones.

- Some representative network covariates:
  - User's past tweeting behavior: the current numbers of total and + tweets.
  - Friends' past tweeting behaviors:
    - ➢ The current numbers of + and - friends (weak).
    - ➢ The current numbers of + and - reciprocated friends (strong).
    - ➢ The current numbers of + and - tweets that these friends have made.

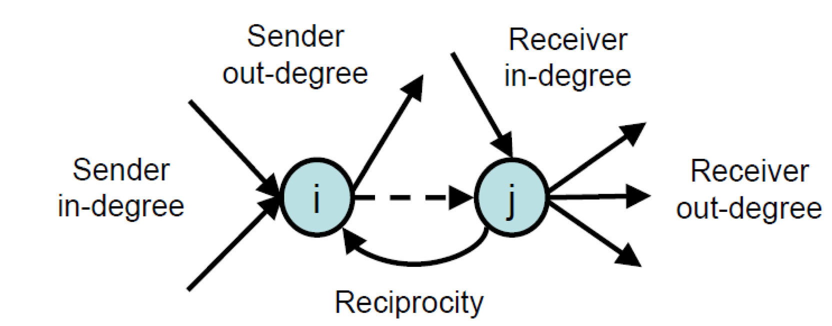
- 95% confidence intervals of coefficient estimates with different sizes of nested controls m:

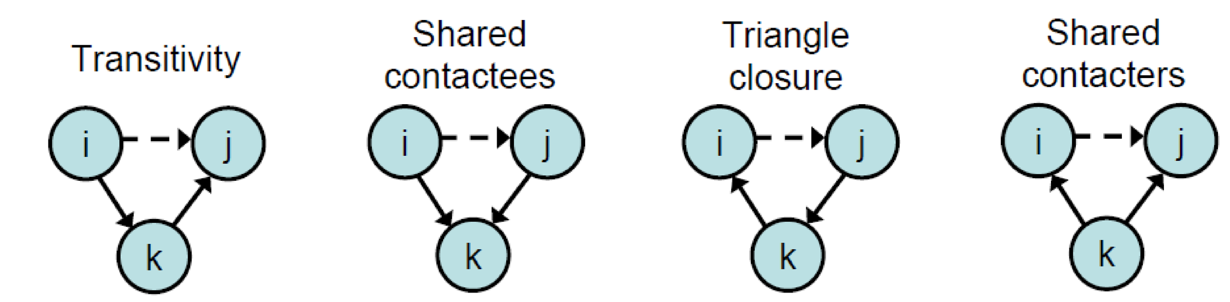


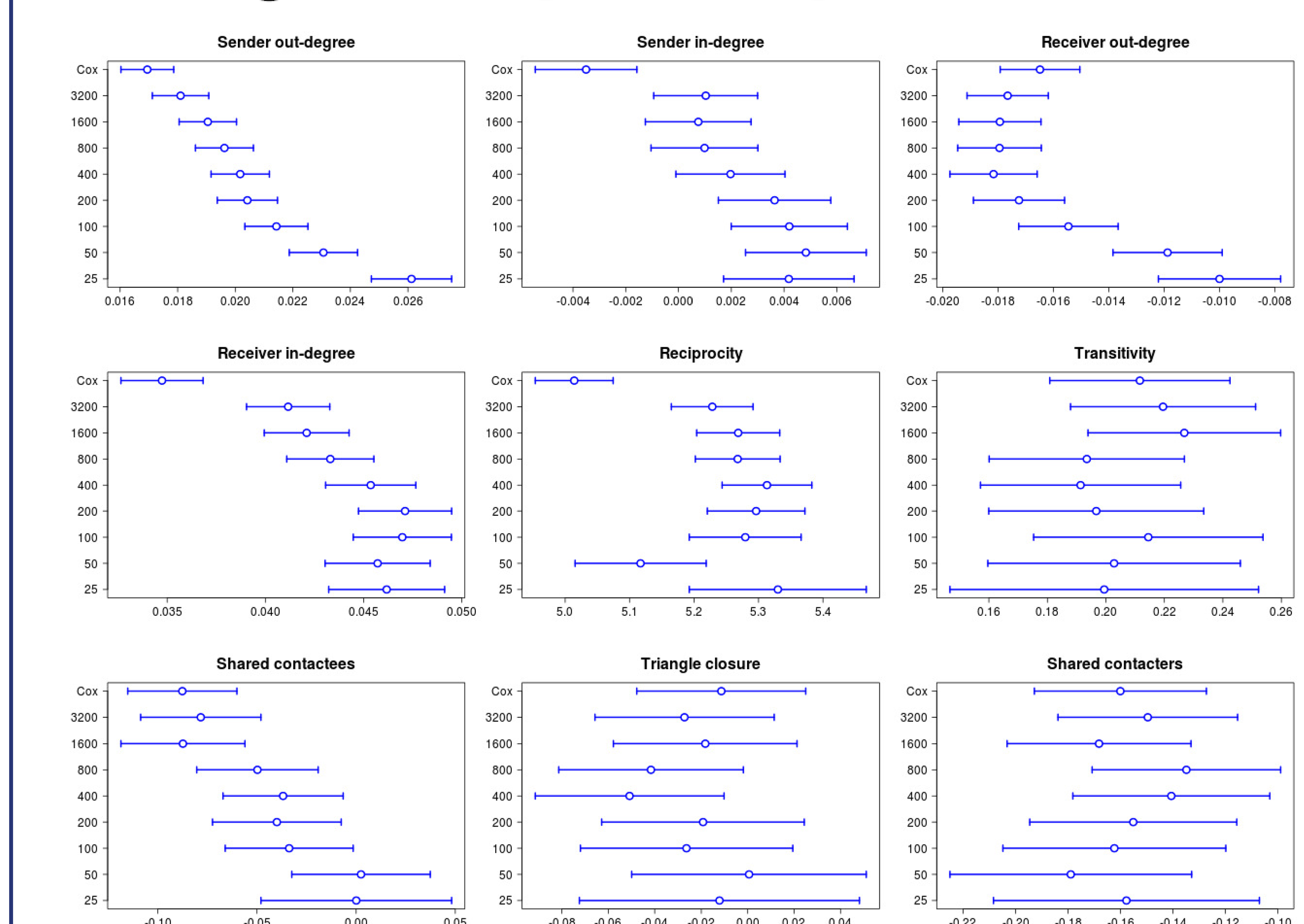
4. Relational Irvine-Facebook Data

- We consider the network formation process of an online social network at UC Irvine [Opsahl et al, 2009] from 5/12/04 to 6/1/04 (7,645 edge formation events among 1,596 active users, i.e. ≈ 2.5 millions of potential edges).

- Network covariates:







5. Conclusions and Future Work

- Nested case-control sampling is simple to implement and fast though biased estimates are possible.

- Other more adaptive risk set sampling methods such as counter-matching will be explored to reduce estimation biases.

- The performance of these sampling methods in the prediction task will also be considered.

- The framework can be also applied to Aalen models with time-varying coefficients [Zhang et al, 1999].