Statistical Learning of Trade Credit Insurance Network Data

Tsz Chai (Samson) Fung

Joint work with Spark C. Tseung and Woongchae (Chae) Yoo

Insurance Data Science Conference

June 20, 2025

M.R. Greenberg School of Risk Science

Introduction

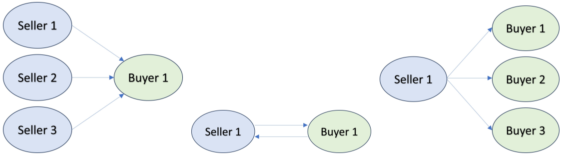
What is Trade Credit Insurance (TCI)?

- TCI is a type of property insurance that **safeguards sellers** against unexpected **risks of losses from transactions** when their **buyer** become **insolvent**.

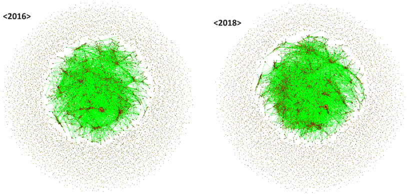
Introduction

What's So Special about TCI?

Dependencies among business entities (sellers and buyers)



Network structure among entities



- **Actuarial literature on TCI data and modeling**: Empty!
- **Goal**: Develop statistical models to **predict** claim probability for each **trade connection** (edge) given network structure.

Data Overview

Proprietary TCI data from a major Asian insurance company from 2015 to 2020:

- **294,272** insured trade connections (network edges)
- **104,494** policies
  - 26.4 % single–buyer
  - 73.6 % multiple–buyer
- **129,915** unique businesses
  - 93,663 as buyers
  - 53,915 as sellers
  - 17,663 in *both* roles
- **6,717** claims in total
  - Binary claim indicator recorded for each trade connection
  - 2.5 % of trade connections have 1 claim
  - 5.66 % of policies have ≥ 1 claim(s)
- **Information captured per connection**
  - Entity profile – status, industry, age, sales
  - Policy details – type, total limit, avg. turnover
  - Buyer-specific limit & turnover

Methodology

Model 1: Logistic Generalized Linear Model (GLM)

- Model claim indicator Z_k for trade connection k through a **logistic regression** given the neighboring information:

$$Z_k | (\boldsymbol{X}_k^B, \boldsymbol{X}_k^S, \boldsymbol{U}_k, \boldsymbol{V}_k) \overset{\text{ind}}{\sim} \text{Bernoulli}(p_k),$$
$$\log \frac{p_k}{1 - p_k} = \alpha_0 + \boldsymbol{\alpha}_1^\top \boldsymbol{X}_k^B + \boldsymbol{\alpha}_2^\top \boldsymbol{X}_k^S + \boldsymbol{\alpha}_3^\top \boldsymbol{U}_k + \boldsymbol{\alpha}_4^\top \boldsymbol{V}_k$$

- Characteristics:
  - Pull information of the associated **buyer** and **seller** as covariates $\boldsymbol{X}_k^B$ and $\boldsymbol{X}_k^S$;
  - Incorporate network characteristics (e.g., degree of centrality) as covariates.

- Limitations:
  - Assume independence of claim index conditioned on observed information;
  - Ignore the influence of network structure, e.g., node importance;

Methodology

Model 2: Generalized Linear Mixed Model (GLMM)

- Model Z_k given some *latent variables* through a **logistic regression**:

$$Z_k|(\mathcal{D}^{\text{full}}, \mathcal{D}^{\text{lat}}) \overset{\mathcal{D}}{=} Z_k|(\mathcal{D}_k^{\text{obs}}, \mathcal{D}_k^{\text{lat}}) \overset{\text{ind}}{\sim} \text{Bernoulli}(p_k),$$

$$\log \frac{p_k}{1-p_k} = \alpha_0 + \boldsymbol{\alpha}_1^\top \boldsymbol{X}_k^B + \boldsymbol{\alpha}_2^\top \boldsymbol{X}_k^S + \boldsymbol{\alpha}_3^\top \boldsymbol{U}_k + \boldsymbol{\alpha}_4^\top \boldsymbol{V}_k + \beta_1 B_k + \beta_2 S_k + \beta_3 P_k$$

- Model *latent variables* B_i, S_i and P_j by **normal distributions**:

$$\begin{pmatrix} B_i \\ S_i \end{pmatrix} \overset{\text{iid}}{\sim} N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad P_j \overset{\text{iid}}{\sim} N(0,1).$$

- Characteristics:
  - Buyer, seller, and policy-level *latent variables* capture network dependence between adjacent trade connections.
- Limitations:
  - Only capture **local** network dependence!
  - Very computationally intensive parameter estimation!

Methodology

Model 3: Network Auto-Logistic Regression Model (NAR)

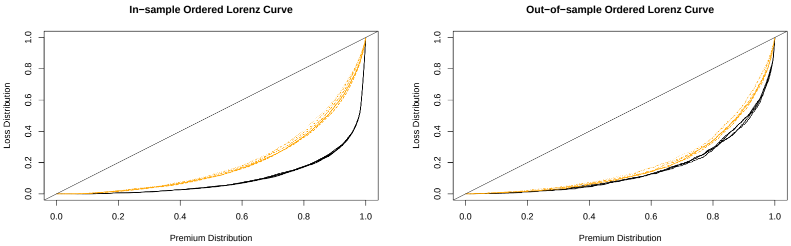
- Model directly the joint claim indicators $\boldsymbol{Z}$ across all trade connections:

$$
\begin{aligned}
P(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Phi}) = \frac{1}{W(\boldsymbol{\Phi})}\exp\Big\{ & \boldsymbol{\beta}\boldsymbol{X}^\top\boldsymbol{Z} + \sum_{l=1}^{L}\gamma^{(l)}\sum_{t=1}^{T}\boldsymbol{Z}_t^\top\boldsymbol{A}_t^{(l)}\boldsymbol{Z}_t + \sum_{l,l'=1}^{L}\delta^{(l,l')}\sum_{t=1}^{T}\boldsymbol{S}_t^{(l,l')\top}\boldsymbol{Z}_t \\
& + \sum_{l,l'=1}^{L}\eta^{(l,l')}\sum_{t=1}^{T}\boldsymbol{Z}_t^\top\boldsymbol{T}_t^{(l,l')}\boldsymbol{Z}_t + \sum_{l=1}^{\tilde{L}}\tilde{\gamma}^{(l)}\sum_{t=2}^{T}\boldsymbol{Z}_t^\top\tilde{\boldsymbol{A}}_t^{(l)}\boldsymbol{Z}_{t-1} \\
& + \sum_{l,l'=1}^{\tilde{L}}\tilde{\delta}^{(l,l')}\sum_{t=2}^{T}\tilde{\boldsymbol{S}}_t^{(l,l')\top}\boldsymbol{Z}_t + \sum_{l,l'=1}^{\tilde{L}}\tilde{\eta}^{(l,l')}\sum_{t=2}^{T}\boldsymbol{Z}_t^\top\tilde{\boldsymbol{T}}_t^{(l,l')}\boldsymbol{Z}_{t-1}\Big\}
\end{aligned}
$$

- Characteristics:
  - Capture beyond local network dependence; e.g., 2-stars effects, 2-step effects, cross-sectional effects, etc.;
  - Fast computation with maximum pseudolikelihood estimation (MPLE).

Data analysis

Prediction results: Ordered Lorenz curve



- Orange: Under GLMs; Black: Under GLMMs/NARs with various settings
- Conclusion: Predictive performance improves by modeling network dependence structure!