Firm-Level Intelligence Tool

Tyche Analytics Company

April 18, 2017

Executive Summary

Tyche was recently engaged to help a carrier evaluate the firm-level characteristics of individual, small firms insured in a large book of program insurance. In particular, we were asked to estimate, for each small firm insured within a program:

(a) the firm's physical location, and

(b) employee headcount

given only:

- the firm's name,
- the occupational class of the workers insured in National Council on Compensation Insurance (NCCI) format, and
- total payroll in US dollars.

Using Bayesian statistics, natural language processing, and open data from various federal agencies, Tyche developed a firm-level intelligence tool that generates probabilistic predictions of firm location and size given only the firm name, occupational class, and total payroll. Using this tool, the carrier can say, for example, that a given policy has a 30% chance of covering employees in an urban area. This knowledge is useful in light of the large disparities in claims experience, where the same claim in an urban area might be 50% more expensive to resolve than in a rural area. Such inferences help carriers to more accurately price program business, where data on policy holders is often quite sparse. This solution is implemented in a web-based tool and can be accessed by browser or API.

Problem Statement

At root, our task is to estimate the posterior joint probability of a location and head count of a program of insured, given knowledge of their NCCI code and total payroll. The firm name is also given, and our very first step in determining firm location and headcount is to search for the firm in the National Provider Identifier (NPI) records provided by the U.S. Department of Health and Human Services. These records contain the information we're looking for—namely, location and firm headcount—but for only a fraction of the firms. The subsequent analysis described below applies to these firms that are not readily accessible in the NPI records.

Materials and Methods

In order to estimate such posterior probabilities, we developed a probabilistic graphical model. To introduce this model, we first describe its elements verbally:

- Each firm has a **l**ocation L, resolved to the county level.
- Each firm has an NA**I**CS class I, where NAICS is the North American Industry Classification System.
- Each worker within the firm has an **N**CCI underwriting class code N, which is related to their **S**tandard Operating Class (SOC) class S.
- Each worker has a **w**age W.
- Each firm has a total **f**irm size F counting all employees.
- Each firm has a **h**eadcount H for workers of given occupational class.
- Each firm has a total pa**y**roll Y for workers of a given class.

The assumed relations between these elements can be summarized as follows:

- NAICS code N is influenced by both:
  - SOC code S—because they are two ways of describing the same information, namely a worker's occupational class
  - location L—because different locations have different concentrations of industry
- Wage W is influenced by location L because of cost-of-living considerations as well as SOC S. We could just as well have chosen to represent this influence through NCCI code N, except for that wage data is only available for SOC codes.
- **F**irm size F is influenced by location L as well as by NAICS industrial category I, which in turn is influenced by SOC code S. Plainly speaking, firm size data is organized by location and NAICS code; thus the SOC codes, which describe *workers*, must be probabilistically converted into industrial codes, which describe *workplaces*.
- **H**eadcount H for a specific class is influenced by firm size F, SOC code S, and industry I.
- Finally, payroll Y is determined naturally by headcount H and wage per worker W.

We summarize these assumed causal relationships in Fig. 1. Here, in keeping with standard notation for Bayesian networks, an arrow from node A to node B indicates that the distribution of A is conditioned by B.[1] Checkmarks denote that data for a node is given, and question marks indicate that a probability distribution over the node is to be derived. All other variables are latent.

[1] See [3] for the standard reference on probabilistic graphical models.

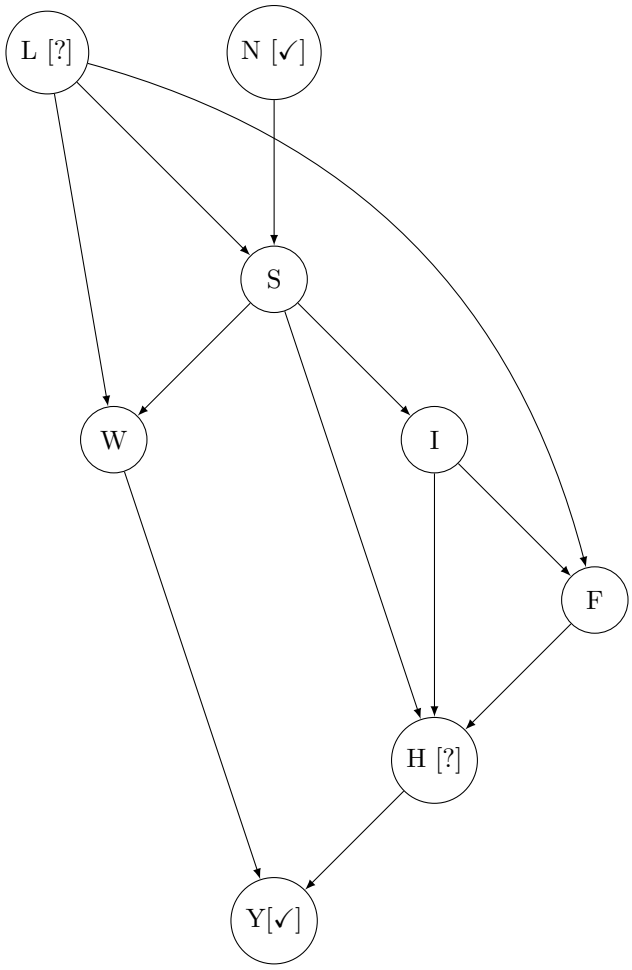

Figure 1: Proposed Bayesian Network of Problem Structure

The arrows in Fig. 1 catalogue the conditional probability distributions we must define:

- $\mathrm{P}(L)$
- $\mathrm{P}(S|L, N)$
- $\mathrm{P}(W|L, S)$
- $\mathrm{P}(I|S)$
- $\mathrm{P}(F|L, I)$
- $\mathrm{P}(H|S, I, F)$
- $\mathrm{P}(Y|W, H)$.

We consider each of these distributions in turn.

$\mathrm{P}(L)$

The prior distribution over location can be derived from a dataset named the County Business Pattern (CBP), curated by the Bureau of Labor Statistics (BLS) [1]. It is simply the probability that a worker chosen at random works at an establishment in a given county.

$\mathrm{P}(S|L,N)$

The distribution of SOC code is assumed to be influenced independently by two sources of information, location and NCCI code. Therefore we may write:

$$\mathrm{P}(S|L,N) \approx \mathrm{P}(S|L)\ \mathrm{P}(S|N)$$

The first term in the right hand side, $\mathrm{P}(S|L)$, can be read off of the CBP dataset. The second term, $\mathrm{P}(S|N)$, is estimated by relating free text NCCI descriptions to free text SOC descriptions with statistical topic classification using Naive Bayes.

$\mathrm{P}(W|L,S)$

Data on the distribution of wages conditional on location and SOC code is also tabulated by BLS's Occupational Employment Statistics (OES) program [4]. For each county-level area and SOC code, BLS reports abbreviated percentile data; in particular we can observe the 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th} percentiles of annual wages. Granting the common assumption that wages within an occupational and geographical bracket are log-normally distributed, we can use this percentile data to extract parameters μ and σ^2 for the wage distribution $W \sim LN(\mu, \sigma^2)$. This gives us a simple parametric form for the conditional probability $\mathrm{P}(W|L,S)$ of the wage W given SOC code S and location L.

$\mathrm{P}(I|S)$

The same OES dataset described in the previous section is cross-referenced by NAICS codes, therefore allowing us to compute the probability that a worker is employed in a certain industry, given that they belong to a certain occupational class.

$\mathrm{P}(F|L,I)$

The CBP also collects data on firm sizes. In particular, the OES data contains binned counts on the number of establishments of type I in location L with certain fixed ranges of employee numbers, e.g. `1-5`, `6-10`, etc. In accordance with established practice in econometric modeling, we assume that firm sizes are log-normally distributed, at least within the range of sizes we consider here.[2] To fit the model, we maximize the log-likelihood:

$$\ell\ell(D|\mu,\sigma) = \sum_{b\in B} d_b \log(\Phi(\log r_b|\mu,\sigma) - \Phi(\log l_b|\mu,\sigma)), \tag{1}$$

where d_b is the number of firms in bin b, and l_b and r_b are its left and right endpoints. Maximum likelihood estimates for μ and σ can then be found numerically.

Shown in Fig. 2 is the fit between observed firm size bin counts for a convenience sample of firms in Augusta County, Alabama, compared to bin counts generated from a log-normal model fit to the same data. Agreement is good.

$\mathrm{P}(H|S,I,F)$

. Headcount can be found by taking the fraction of workers with SOC code S in industrial code I and multiplying by firm size F.

[2] Agreement with the log-normal model breaks down in the right tail, where a power-law distribution is more appropriate.

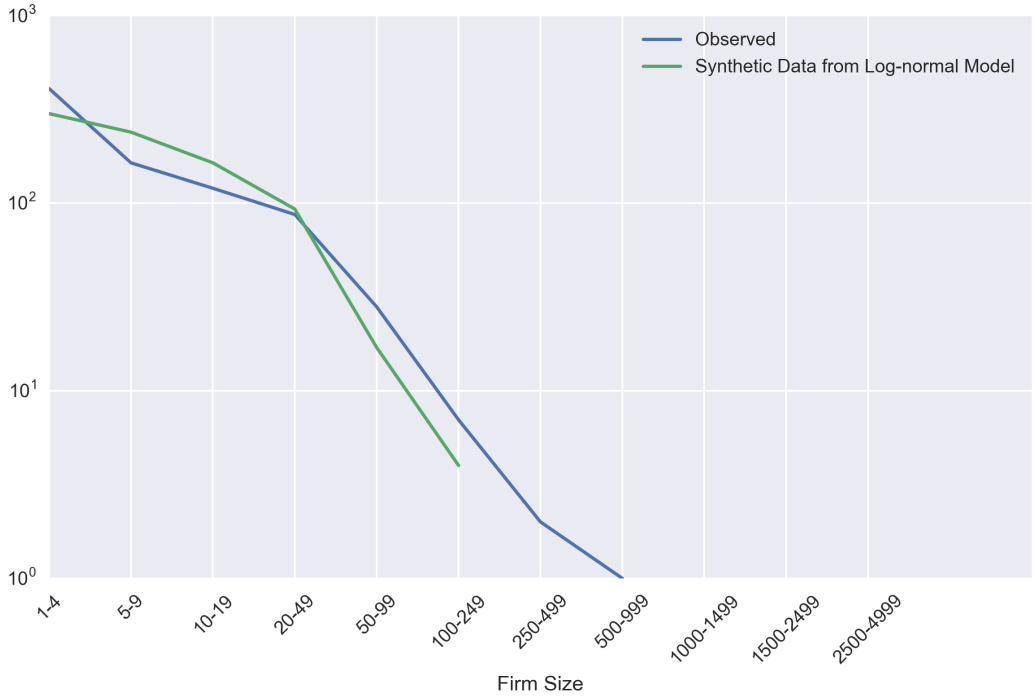

Figure 2: Comparison of observed to log-normal synthetic firm size data

$\mathrm{P}(Y|W,H)$

. Naturally, payroll is related to wages and headcount through the simple equation,

$$Y \sim \sum_{i=1}^{H} w_i, \tag{2}$$

where $\{w_i\}_{i=1}^{H}$ are realizations of the wage variable W. Unfortunately it is one of the scandals of statistics that the distribution of Y, being a sum of log-normal random variates, is almost totally opaque to analysis [2]. In light of this difficulty we can proceed either by Monte Carlo simulation, which is theoretically exact in the limit of infinite sampling but introduces simulation error for finite sample size; or a moment-matching approximation, which is theoretically inexact but immediate.

For ease of analytical tractability, we opt for the latter approach. If $W_i \overset{i.i.d.}{\sim} LN(\mu, \sigma^2)$, then the mean m and variance s^2 of W are given by,

$$m = e^{\mu + \frac{\sigma^2}{2}}, \text{and}$$
$$s^2 = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2},$$

and from Eq. (2) we may match moments and write,

$$\mathrm{E}[Y] = Hm, \text{and} \tag{3}$$
$$\mathrm{V}[Y] = Hs^2, \tag{4}$$

Given the first two moments of Y in Eqs. (3, 4), how should we model the distribution? Two arguments speak in favor of assuming a normal distribution: the central limit theorem, which asymptotically governs all sums of independent r.v.'s with finite variance; and the principle of maximum entropy, which recommends the normal distribution for modeling random variables subject to the constraints of known mean and variance.[3] Thus we can write,

$$\mathrm{P}(Y|W,H) \approx \phi(P|Hm, Hs^2), \tag{5}$$

where ϕ is the normal p.d.f.

Implementation and Results

The model is implemented in Python and exposed via a Flask web app, allowing the user to input queries in the form of occupation and payroll data and receive probabilistic estimates of location and headcount, visualized as choropleths and histograms respectively. API integration is also possible. Sample output is shown below:

[3]Strictly speaking, payroll is also subject to the constraint of non-negativity (one hopes!), but this constraint is negligible in practice.

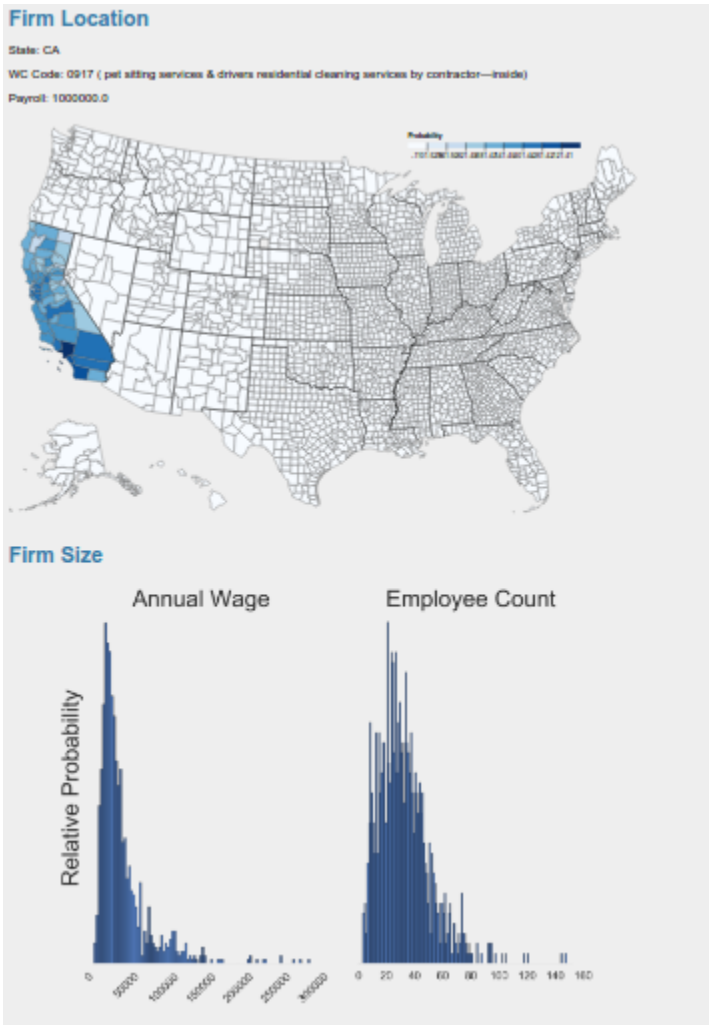

Figure 3: Sample output of the firm-level intelligence tool, showing posterior distributions over location (above) and annual wage and headcount (below)

References

[1] U.S. Census Bureau. County business patterns, 2014. Accessed at `https://www.census.gov/data/datasets/2014/econ/cbp/2014-cbp.html`.

[2] D. Dufresne. Sums of lognormals. In *Proceedings of the 43rd Actuarial Research Conference*. University of Regina, 2008.

[3] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[4] Bureau of Labor Statistics. Occupational employment statistics, 2016. Accessed at `https://www.bls.gov/oes/oes_ques.htm`.