Technical Report: Accessing Labor Data on Wage Theft Judgments

Info

This document serves as an incremental writeup.

See

- https://docs.google.com/document/d/1XJLxRPccDk96_LJwPNQir8RvYkg9SlZrLcz8kU-eZ2g/edit
- https://docs.google.com/document/d/1iP6E1u7zB7TGQbbM3kSlZ9OKotb_YEFVrVzOfD1onRM/edit#heading=h.9i1r03ryd3yp

Meeting schedule

- Weekly technical update meeting
  - 2 hours
  - In-person
- Monthly substantive strategy update
  - Center for Industry Facility Engineering (Forest and Martin)
  - Santa Clara County Wage Theft Coalition (Michael T. and Ruth)
  - Foundation for Fair Contracting (Arturo)
  - Department of Labor Wage and Hour (Michael)

We have had seven meetings to date (late Juy) at Stanford, San Francisco DataKind, and the Santa Clara Wage Theft Coalition

Action Items

Next milestone

- DataKind DataDive (Fri, Sat, Sun) weekend of July 29th or August 26th
- Turn in grant proposal (mid-August)
- Nate begins research assistantship (Fall Qtr)

Housekeeping

We should use a task sharing app like asana

Action Items for two weeks ending 8/5/2016

- Santa Clara violations by zipcode; handoff to Michael.T early August
  - Make a dive into the restaurant industry in the Santa Clara region; remove SCA cases from the dive (talk with Ruth or Michael T.)

  - Create visualizations in the next month by city for the south bay area to support the efforts with city councils to enact wage theft ordinances
- Prepre for DataKind DataDive end of August
- data sources tie together
- need to haves for data
- concerns with data; example: span of years
- share technical report of findings (journalists Caroline or Ruth)
- Develop a validation plan for our predictive models
- Begin discussing the use of the DOL API
- Discuss the cross of the DOL headers with the CA Labor Commissioner headers

Action Items for two weeks ending 7/21/2016

If any of these items sound interesting - reach out to the listed person (7/8)/

- Ash: write up a short report on his ROI index
- Forest: write up a short report on the DOL dataset headers (done 7/3)
- Thor: look at future use cases

Open action items looking for a volunteer:

- Standardize data sources and map together
- Look at the API on the BLS site
- Cross the DOL dataset with the California Labor Commission dataset

Overall, continue to look at low wage occupations and minimum wage violations. As a case to develop and validate the formalizations, let's use the Santa Clara County (San Jose Metro) data - both DOL and California Labor Commission.

Action Item for week ending 6/10/2016

- Replicate the detective work the DOL does with the data to identify trends such as the agricultural crop with the most complaints (in-progress)
- Onboard DataKind volunteers (Done 7/29/2016)
- Collect datasets from Foundation for Fair Contracting (In progress)

Progress Summary

The last meeting was a strategic meeting with the Santa Clara Wage Theft Coalition where we picked up some key action items.

- Create visualizations in the next month by city for the south bay area to support the efforts with city councils to enact wage theft ordinances

- Develop a validation plan for our predictive models
- Make a dive into the restaurant industry in the Santa Clara region; remove SCA cases from the dive
- Begin discussing the use of the DOL API
- Discuss the cross of the DOL headers with the CA Labor Commissioner headers

We are rounding a key milestone where we have become comfortable with the data through our individual dives and when we talk about our dives we are beginning to have a common understanding of the data. We are going to start ramping up for the next DataDive with DataKind.

Dataset List

- WHD Investigations and Violations by zipcode and NAICS occupation2002 -2015
- San Jose Office of the District Attorney: Judgments 2014 – 2016
- San Francisco Department of Labor cases
- BLS: population and wages by SOC occupation and BLS metro region
- Lookup Datasets
    - FIPS State Codes
    - BLS Metro Regions
    - Zipcode database
    - SOC structure
    - NAICS structure

Pilot

Functions and Calculators

“Enforcement Payoff” (Ash) score provides a metric for the industries “flying under the radar.” This is working off the DLSE WHD Enforcement dataset **only**. It surfaces industries where there is a consistently large payoff (understood in terms of Total Violations Discovered) given at least a few enforcement cases in the dataset.

Dives

San Jose Metro Region Construction Industry (Forest)

Data Analysis Standards

Equalize DOL time series with BLS instance demographic with the mean DOL compared with the BLS instance

DataKindSF One-page Summary

What are the Wage Theft Coalition and the Wage and Hour Division of the Department of Labor?

The Wage Theft Coalition is a collection of community organizations in Santa Clara County (aka Silicon Valley); their mission is to stop wage theft.

The Department of Labor enforces the Fair Labor Standards Act (FLSA), which sets basic minimum wage and overtime pay standards; the Wage and Hour Division enforces these standards.

What is the biggest problem these organizations are currently facing?

The goal is to stop the predatory exploitation of vulnerable populations.

Wage theft is the practice of: paying less than minimum wage, failing to pay overtime, forcing work off the clock, issuing paychecks that bounce, stealing tips, denying required meal and rest breaks, misclassifying work (i.e., as independent contract work), and not paying at all. Even when a worker exercises their right, they face illegal retaliation. Even with a court award, it is hard to collect; in 2012-13 workers collected 20 cents on the dollar.

The Wage Theft Coalition needs to show public policy makers that wage theft is a problem in their jurisdiction and that they must pass ordinances to end exploitation.

Wage theft continues despite investigation and enforcement by the Department of Labor. They have the classic ‘missing species’ problem: The violators they have discovered are known but the violators they have not discovered are unknown. Predictions using the existing data simply predict known violators. The goal is to find the gaps in the dataset where violators are ‘flying under the radar.’

What data do they have so far?

- Dataset compiled from investigation and complaint source cases; organized by NAICS and unique ID that covers a case. Each case is an event of a single company, multiple employees, and multiple violations. Each event has a new unique ID.
- Dataset with field for categorizing source of case (not available to public)
- Datasets of industry information from other departments such as OSHA, MSHA, Census, Dept of Commerce, ACS
- Datasets from advocate groups enforcement such as Building Trades Council, Foundation for Fair Contracting, and the California Labor Commission

What have they done with the data so far?

- Basic online data visualization
- Online csv download
- Data analysis such as on agricultural crops with highest complaints
- Integration of datasets from various federal departments (discussing API)
- Developed app ‘Eat, Shop, Sleep’ of violators

Key questions we’d like to discuss tonight

1. Identify the features of where wage theft impacts vulnerable populations
2. Formalize estimating to what degree wage theft impacts a population
3. In what way to communicate/visualize the dataset so advocates can A) draw conclusions and B) host a discussion

Mandate

The Wage Theft Coalition needs to close loopholes in the law: The County of Santa Clara needs an ordinance that allows revoking city permits and privileges (restaurant and similar) for unpaid wage theft judgments. Currently, the county revokes the contracts with companies although there are a lot of companies seeking waivers through a loophole.

The goal is a platform for combining nationwide all the datasets on wage theft from federal, state, city, and private entities. Today, this platform does not exist. However, the federal nationwide data is available, and there are datasets maintained by states and private organizations on wage theft cases. This platform will be helpful for both advocates and workers.

The statisticians will have a Phase A with the DOL data and after piloting the DOL data, there will be the DLSE in a concurrent phase B.

We would like to involve Julie Su who heads up the DLSE. She was a former advocate who brought the garment industry cases and labor trafficking cases in L.A. The Wage Theft Coalition (the Coalition) is interested to have the DataKind volunteers look at the Labor Commissioner (DLSE) data because most of the Coalition clients go to the Labor Commissioner because state law is better in this case than federal law. The Labor Commissioner dataset was obtained through a Public Records Act request and are in the shared Box cloud folder with the Department of Labor (DOL) data. Much of the state data can be made publicly available through FOIA requests, some states may already be publishing their data publically, and other states may be willing to move in that direction.

We need to look at the DOL because they have enforcement actions at businesses and collect a lot of data that is not just generated by individual claims. It will also be very helpful, and Michael Eastwood is very motivated to use our project data for the good.

The goal is a one stop platform for information on specific companies, aggregate data on cities, counties, and states, and (to the extent that the other datasets include NAICS codes or equivalents) data on industries. This could become a center for research, with forums/groups where you discuss different uses of the data, posting of reports, etc.

Introduction

The Department of Labor has the “missing species problem.” The DOL has a dataset of the known wage violators and from this they can find the type of violators and their domains. The problem, we know there are unknown violators and unknown types of violators that operate in unknown domains. Finding the unknown from the known is challenging.

With this, the case can be made to each city council to provide ordinances that undercut the employers that are at the core of exploitation.

There are pragmatic metrics that have become clear over time to those who investigate labor violations. The key indicator is an exploitable population. The more exploitable the more likely they are being exploited. This means we need a way to measure each population’s exploitability. A second metric is the lack of violations. The lack of violations is not an indicator of a lack of exploitation it is a measure of a domain that is running under-the-radar. We need to know the domains, the expected rate of violations in each domain, and then the actual degree of violations in each domain. The third metric is low wages. Where there are low wages there are exploited workers that cannot defend themselves. Where these three metrics unify is where the department of labor pragmatically has found is the best place to investigate. At the core a domain with low wages and no complaints means there are likely violators.

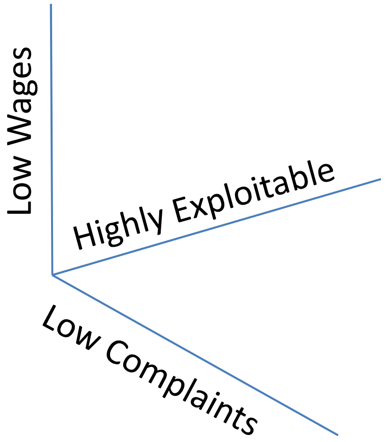

Figure 1From the pragmatic experience of the Dept of Labor investigators, the three core signs of exploitation is low wages, low complaints, and an exploitable population.

The consensus is that regardless the solution to this problem, the solution needs an API interface to prevent the underlying data from becoming stale.

An ancillary concept we must be aware of is that lack of complaints in the Dept of Labor dataset does not mean no complaints. It could be that is not a target domain for the DOL. There are other groups operating in the wage theft domain such as the California Labor Commission and the Foundation for Fair Contracting. The Dept of Labor has only a few construction related violations but this is not from a lack of complaints. The Foundation for Fair Contracting has been active in enforcing this domain and likely the violations are with the California Labor Commission since the California laws are more favorable than the Federal laws offer. This case shows the need to fuse/integrate/mash multiple datasets to get the full domain of data.

Our audience in this project is the investigation and enforcement agency decision makers, the elected legislative policy makers, the press, community advocates, and the public. The big goal is to host a platform for discussion amongst these groups so to end the exploitation of those least able to protect themselves.

Wage theft Project Breakdown

Structure

- Product
    - Action
        - Discipline
            - Operation

Wage Theft Project

- Methodology: platform for discussion
    - Retrospective study
        - Data Engineer
            - Acquiring
                - FFC dataset
                - BTC dataset
                - BLC dataset
            - Ingesting
            - Transforming
            - Storing
            - Retrieving data
                - API
            - Meta data tags
                - Convert NAICS to SOC format
        - Scientist (Scientific method)
        - Mathematician
        - Statistician
        - Engineer (Advanced Computing): development from specification
        - UI/UX: Visualization
        - Domain Expert: Subject Matter Expert
    - Predictive formalization from features

The Datasets

Violation Datasets

- San Jose Office of the District Attorney: Judgments 2014 – 2016
- San Francisco Department of Labor cases
- DOL-WHD Investigations and Violations by zipcode and NAICS occupation2002 -2015
    - http://ogesdw.dol.gov/views/data_catalogs.php
    - See USB and emailed link to excel
  - Feature for complaints versus investigation (feature not public)
    - Probably something public online that indicates the percentage that are complaints versus investigation originated
  - How is data collected (features)
    - Walk in, complaint, or call in – collect feature data
      - Location
      - Industry (NAIS)
      - Company name, address, phone
      - How heard about at DOL
  - Fields
    - Violation type
    - # of employees
    - Unique ID for each case: Case is a bundle of employee/company/violation
  - Decision to pursue case is to follow those least able to help themselves
  - Complaint screenings for vulnerability
  - oesm15all.zip from DOL website http://www.bls.gov/oes/tables.htm May 2015 all data
- California Labor Commission (Ann) Department of Labor Standards Enforcement (DLSE) obtained by Ruth through Public Records Act Request – data is of recorded a judgments in Santa Clara Superior Court
    - 1) April 2011 to April 2014
    - 2) April 2014 to January 2016
- Building Trade Council (BTC) (Josue)
- Foundation for Fair Contracting (FFC) (Jesse)

Lookup Datasets

Integration of industry information (Kris advises this is tricky/infeasible)

- Good to have:
    - Current DOL resource allocation to normalize dataset
    - DOL complaint versus field enforcement dataset
- Bureau of Labor Statistics (BLS): population and wages by SOC occupation and BLS metro region
- SOC structure
- NAICS structure
- OSHA

- MSHA
- Census Employment data by industry: County Business Patterns (U.S. Census) https://www.census.gov/econ/cbp/download/ It has number of establishments, employees, and payroll information by NAICS. In the attached file: data by NAICS code at the 3 digit level with enforcement data on violation cases from the Wage and Hour public data for the counties that are within the jurisdiction of the San Francisco District Office (Del Norte, Humboldt, Mendocino, Sonoma, Napa, Marin, Contra Costa, Alameda, San Francisco, Santa Clara, Santa Cruz, Monterey, and San Benito).
- Dept of Commerce
- American Community Survey (ACS)
- FIPS State Codes
- Zipcode database
- https://github.com/bokeh/datashader
- Looking for
    - undocumented workers by industry and metro region
    - demographic by NAICS instead of SOC

Research Plan Outline

Big goal

Stop the predatory exploitation of vulnerable populations through technology assistance to the enforcers of the current laws.

- Useful internally to DOL (under-the-radar domains/industries for DOL resource allocation/ who should be investigated next)
- Assessable to advocates to invoke in campaigns (example, Ruth's city ordinances)
- Predictive model
- Create a discussion (press)

Problem

- Currently, the WHD decides what industries to investigate for possible wage theft violations based on more or less anecdotal evidence. It would help the WHD allocate resources if there were a systematic way of using the data they collect to make these decisions.
- The Department of Labor has invested in visualizing some of the data it collects, but so far they haven't put much effort into the wage theft data (they also don't seem to have anyone actively working on this). Having an interactive visualization that allowed users to query wage theft data and interpret it in its appropriate context would be useful both internally at the WHD and externally as a resource for advocates campaigning around worker's rights.

Intuition

- Linking to industry information might be tricky
- Problem industry
  - Problematic industries (by NCIS code)
  - Similar industries (A is problematic so B should be investigated)
  - Specific problematic businesses
- Problem geography regions
- Domains with (gap) non-utilization of DOL (regional/industry/population)
- Emerging problems
- Large problems
- At risk population demographic
  - Low wage domains
  - Immigrant
  - Education
- Define by
  - City (zipcode)
  - Industry (NAICS code)
- Industries with characteristics to avoid due to difficulty enforcing
  - No industry hierarchy
  - Businesses open/close velocity high

    - For example: Restaurants will show but industry of this type difficult to enforce so avoid
- Residential Care Homes (NCIS 623) should show (validation metric)

Past attempts

App

- (failed for various reasons): 'Eat, Shop, Sleep' app of company violators to avoid using these companies
- Prof Fischer suggested 'Look Wisdom' energy assessment app
- MSHA visualization http://ogesdw.dol.gov/homePage.php

Analysis projects

- Online website visualization
- Link to available csv file download
- Analysis for which agg crops have most complaints
- Integrated data (fusion)

Questions

Stated Null Hypothesis

How best is the public discussion hosted → Big Question

Framing actionable problem statements to solve (AM)

1. Provide the DOL WHD with some way to get an idea of "similarity" of violation patterns between NAICS sectors
   1.1. At a national level
   1.2. At state and lower levels
2. Provide a visualization of the WHD dataset, where the user may browse the map, zoom in/out of geographic area and view details of individual violations from the dataset.
   2.1. What visualizations satisfy the needs of advocates and is assessable "On this note, it would be valuable to be able to adapt this hypothetical interface to subsets of the data, because some advocates might care mostly about specific industries or communities (e.g., home health or recent immigrants). The primary challenges in this project are identifying what views of the existing data are most informative and determining what supplemental data should be incorporated."
3. Provide a way to identify "at-risk" populations by tying demographic data to the WHD dataset. This possibly is an enhancement to 1.1 and 1.2 above.
   3.1.1. Where does wage theft currently impact the at-risk population (who should be investigated next) "For this project, one of the main challenges will be to pinpoint what exactly is meant by an "interesting" industry. The WHD already knows there are some industries where exploitation is common (in fact, this influences the industries that show up in the data); also it wants to maintain a focus on the workers who are most vulnerable."
   3.1.2. How is wage theft currently impacting the at-risk population (where is ROI)

3.1.3. Where is wage theft predicted to impact at-risk populations

Specific analysis goals

- Lack of enforcement (under-the-radar domains)
- Population vulnerable to exploitation (wage-education-migrant)
- High impact measured in individual case dollars (example, less than $7.50 min wage violation/low wage industries)
- Ease of enforcement (ex. restaurants are too difficult to enforce)
- Predictive
  - Predictive factors
  - Predictive model

Tasks

- Initial Pilot: Focus analysis on wage theft in Bay Area – construction industry as a first pilot
- Distance measurements
  - Measure Degree of Population Exploitability
  - Measure Expected Industry Degree of Violations
  - Measure the Occupation Wage (from Labor Stat API)
- Normalization
- Analysis
  - Rank city by wage theft
  - Rank city by backwages
- Visualizations
- Model-based predictions
  - Define factors
  - Define factor significance
  - Predictive model
  - Learning
- Updatable final product (API integration)
- Types of data analysis: undefined
- Organizing: undefined

Validation

Cross validation

Contribution

- Analysis standards and metrics for wage theft
- Framework for predicting wage theft; define new features that improve the ability to answer
- Formalization for predicting wage theft → tools/dashboard/API specification

Impact

- Increase in wage theft enforcement return on investment

- Analysis standard available for use by other regions
- Sharing of wage theft information for policy and enforcement
    - Press release with collaborators (press strategy)
    - Potential end users
        - Media
        - Foundation for Fair Contracting
        - Building Trades Council
        - Dept of Labor
        - Cal Labor Commission
        - Wage Theft Coalition

Suggestions for future work

If this goes well there is a possible phase II to look at the effectiveness'of enforcement tools crossed with the resource intensity of the enforcement tool (ROI) by intervention tool and type of intervention; this could be an officially sponsored DOL project.

Preliminary Case Studies

Notes

Functions and Calculators

"Enforcement Payoff" (Ash) score provide a metric for the industries "flying under the radar." This is working off the DLSE WHD Enforcement dataset **only**. It surfaces industries where there is a consistently large payoff (understood in terms of Total Violations Discovered) given at least a few enforcement cases in the dataset.

Data Analysis Standards

Equalize DOL time series with BLS instance demographic with the mean DOL compared with the BLS instance

Task List

Framing actionable problem statements to solve (AM)

4. Provide the DOL WHD with some way to get an idea of "similarity" of violation patterns between NAICS sectors
   4.1. At a national level
   4.2. At state and lower levels
5. Provide a visualization of the WHD dataset, where the user may browse the map, zoom in/out of geographic area and view details of individual violations from the dataset.
6. Provide a way to identify "at-risk" populations by tying demographic data to the WHD dataset. This possibly is an enhancement to 1.1 and 1.2 above.

To keep things manageable to start with, I'm focusing on **1.** for right now. *I'm basically going to evaluate the "similarity" between overall violation patterns in NAICS sectors nationwide and provide some sort of a graph of other NAICS sectors that have similar offending patterns to a particular sector.* My hope is that this serves as a "recommender" system for the DOL WHD in knowing which other sectors to tackle next and what to expect based on similarity to a given sector. Part of my attempts in determining "similarity" here will include trying to surface sectors where not as many violations are reported but where this isn't reflective of reality, i.e. the nature of violations and the proportion of workers affected are more severe. I'm basically trying to solve (or at least prototype solutions for) at least a couple of the problem statements in the next couple of weeks.

On the visualization side it would be great to be able to zoom in to a state, county, and city level, and at each level be able to see/access data on:

1) Specific companies with violations, ranked by size of back wages.
2) Data by industry, i.e. industries with the most violations, most back wages, etc.. Also ability to filter by industry. Probably need high level industry (2 digit NAICS code) and then also ability to drill down to more detailed industry level, (maybe 3, 4, or 5 digits).

3) Geographic data, i.e. dots for each case where user can click to bring up information on that particular case. This would be especially true of the county and city level, but also maybe the state level in smaller states. This could also be done as a heat-map.
4) Filter by law, i.e. only see Minimum Wage and Overtime cases (Fair Labor Standards Act), Government Contract cases (Davis Bacon and Service Contract Act), agriculture violations (Migrant Seasonal Worker Protection Act and H-2A Act + Fair Labor Standards Act with agricultural NAICS code),
5) Ability to create links and embedable code to specific filtered views, so that an advocacy group that is working on advocating for a particular industry, say home care workers in a certain geographic area, could use the visualization tool to create a map showing all cases in that industry in that area, and then create a link to that map and/or code to embed that map on their web site.

This is harder, but it would also be great to be able to add in supplemental datasets (from states, non-profits, etc.) that could also be visualized at the same time. Probably the number of available filters and analysis (like by industry) would have to be seriously restricted when adding in other datasets, but if the other datasets have industry information, you could still use that, and if they have back wage amounts, you could still do case ranking and geographic analysis/plotting.

Nationwide and local (i.e. state and county level):

1) What industries have the most severe average minimum wage violations (i.e highest average BW/person number)? - This is a measure of severity.
2) What industries have the largest number of workers paid less than MW? - This is a measure of how many workers are impacted.
3) What geographic areas have the highest severity cases (#1 above) and most workers impacted (#2) on an absolute level and relative to population (both numbers are useful for planning purposes)
4) Is there a relationship between #1 and #2 and census data? What is the best predictor for #1 and #2? Population size will obviously be correlated with #2 but what else? % of population that speaks a foreign language? % of population that is not white? Does the size of certain ethnic groups correlate with #1 and #2?
5) What are the main differences between region, i.e. between states and counties? Are there specific areas where industry A has severe violations with large numbers of workers impacted but in other areas the violations area severe but very few cases are done? This would be great as it would help with identifying industries where one region has been turned on to a problem but others have not.
6) BLS/Census data: What industries have the lowest paid workers? What are the industries in a given area that have the most number of workers who are low paid and who speak English as a second language?
7) What is the relationship to the industries identified in #6 above and the WHD data? Are there any industries that based on #4 and #6 we would expect there to be lots of problems, but there

is not a lot of cases? And if so, are there any industries where we would expect to see problems and there are not a lot of cases but when there are cases there are severe violations?

8) Simpler variation of question 7: What industries have severe violations, a large number of workers (based on BLS/census data), but not many WHD cases? (i.e. cases are severe when done, but not many are done)?

For purposes of the above, it probably would be best to exclude any cases that have violations listed that are not Fair Labor Standards Act related. That way you are limiting your universe to cases that are about MW and OT.

Later, we could add in analysis like of government contract cases and other Acts but for now I'd keep it simple.

Crossing Frameworks

AM: After our conversation on Thursday I began looking at the main WHD dataset up at the DOL Enforcement website (Wage and Hour Compliance Data).

I've spent quite a bit of time hunting down other data sources to get an idea of overall employment numbers by NAICS codes and even demographic information. The trouble is that, between the numerous Occupational Employment Statistics (OES) and Statistics for US Businesses (SUSB) data sources, the NAICS codes are simply not consistent as they have evolved over time. I managed to account for around 86% of NAICS codes in the WHD data in the SUSB data, and this helped me get a *national* level idea of overall employment numbers by NAICS sector.

Demographic information is next, and there are some other sources at the census portal that might have something. Similar issues exist with tying demographic information by NAICS codes, by region, from numerous fragmented and often unclear data sources. But I'm convinced that it **is** a feasible task.

FP: This afternoon I looked at crossing the NAICS codes with the SOC codes. They don't overlap verywell.

NAICS is industry segment based (top down) and the SOC is task based (bottom up). I mapped between them for the construction industry.

There must be a standard somewhere for how these relates. One of the datasets had both NAICS and SOC colums but the NAICS codes were all 000000 all industries. I took that to mean this is problem.

AM: Good point. I too noticed that the data that was available at the OES (BLS site) in this case was broken down by "Occupational Code" (I think SOC), in addition to the NAICS not overlapping very well (only a 2% coverage with the 2013 OES data). It would be really helpful if the WHD data had affected occupational code as additional info, going forward.

FP: I agree, the codes are at differing levels of detail. The NAICS is an object type level like build bridge and build highway. The SOC is skill level like ironworker or cement mason. The codes go together as NAICA-SOC. You have more complete context of the company type and worker skill such as bridge.construction-ironworker.

Those are great questions and totally on point! The NAICS codes are a bit tricky for sure. One approach that I've used is to not try and go too deep, i.e. cut off the codes at 4 digits or 5 digits. One thing you'll notice is that in any individual case, there might only be 4 or 5 digits of detail entered anyways - it is not a requirement for each case to have the full NAICS code. In other words if 10 is agriculture, and 101 is row crops, and 1012 is legumes, and 10122 is broccoli, some cases involving broccoli might have 10122 selected as the NAICS code and others might only have 1012 selected.

So, if you are trying to do your analysis at the 5th digit level of detail, you will miss all the cases where only 4 digits were entered. In addition, for purposes of analysis, if you do the full 6 or 7 digits (I forget how many digits a full NAICS code is), you will end up with a ton of meaningless sub-categories, so sticking with a somewhat higher level grouping is probably good just from a data analysis perspective anyway. It's trivial to do that - just add a column with only the first 4 or 5 digits.

That will also help a lot with matching NAICS codes to other data sets.

Test Dive into South Bay Construction

San Jose Metro Region Construction Industry (Forest)

Cases with no employees or assessed penalty

30% of the San Jose Metro cases has no employees and no assessed penalties. I am working on the assumption that

- No penalty: these are complaints that resulted in a finding of 'No Violation' found
- No employees: these are undocumented workers

NAICS cross with SOC

Case drive on San Jose Metro Construction Industry (NAICS 23) due to my familiarity

There are NAISCS cross with SOC on BLS site (cannot find excel file) http://www.bls.gov/oes/current/naics5_238140.htm The NAISCxSOC example for "NAICS 238140 - Masonry Contractors" needs improvement.

Thirteen (of 42codes) NAISC codes map to multiple SOC codes – but in practice the DOL uses these thirteen codes for 55% off all cases; see next section on 'other' codes

- 236000 Residential and Commercial Building, (maps to 45 crafts)
- 237200 Heavy:Land Subdivision (only 3 cases, 0% of total)
- 237300 Heavy:Highway, Street, and Bridge (maps to crafts: Laborers, Carpenters, Operators, Masons, Teamsters, Ironworkers)
- Other
    - 237900 Heavy:Other (see other codes)
    - 238190 Specialty:Other Foundation, Structure, and Building Exterior (see other codes)
    - 238290 Specialty:Other Building Equipment (see other codes)
    - 238390 Specialty:Other Building Finishing (see other codes)
    - 238990 Specialty:Other Specialty Trade (see other codes)

53% of codes placed in an 'other' category and based on a sample audit of San Jose Metro 75% of 'other' codes are 'code dumping'

- 238190 NAICS: Specialty:Other Foundation, Structure, and Building Exterior
    - San Benito Heating and Sheet Metal, Inc. →miscode (238220:MEP)
    - D.H. Smith Company, Inc. (drywall) →miscode (238310:drywall)
    - Vickers Concrete Sawing, Inc. → correct code
    - Jensen Landscaping → miscode(561730:landscaping)
- 238900 NAICS: Specialty:Other Specialty Trade
    - 11 cases in various sub codes
    - Looks like dumping
- Suspect that "Building: Residential & Commercial" is used like an 'other' category

Conclusion: 40% of construction activities are 'dumped' into 'other' and are of no clear trade

Looking at the entire state of California 55% of cases in an 'other' category, therefore this pattern looks to hold true generally from what I saw in the San Jose Metro

My intuition is the 'other' code categories are capturing workers that have no defined trade and are multicraft. They may be carpenters one day, plumbers the next, and then work as electricians. This role is tyical of non-union workers and makes categorizing the worker difficult. Some specialty trades will have a defined type of work they typically bid and these most likely account for the other 45% of workers. Zeroing on the workers that transit between occupations as a part of their occupation is a class of worker in itself.

Bad guys are just bad guys at this point: All across the board will be compliance cases, no violations and Could be any Act, family leave, H-1B, etc. I was just looking at a ballpark if there was four or twelve - I started with construction because I know the industry well enough that I can tell if the charts don't make sense. Once I get the process down then I can replicate for all the other industries and assume it worked for construction then the results should be valid for the others. I think the main takeaway is there is a huge 'other' category to be looking for when we move over to care homes or restaurants.

To validate this intuition I suggest interviewing a sample of violators and ask about the workers craft category.

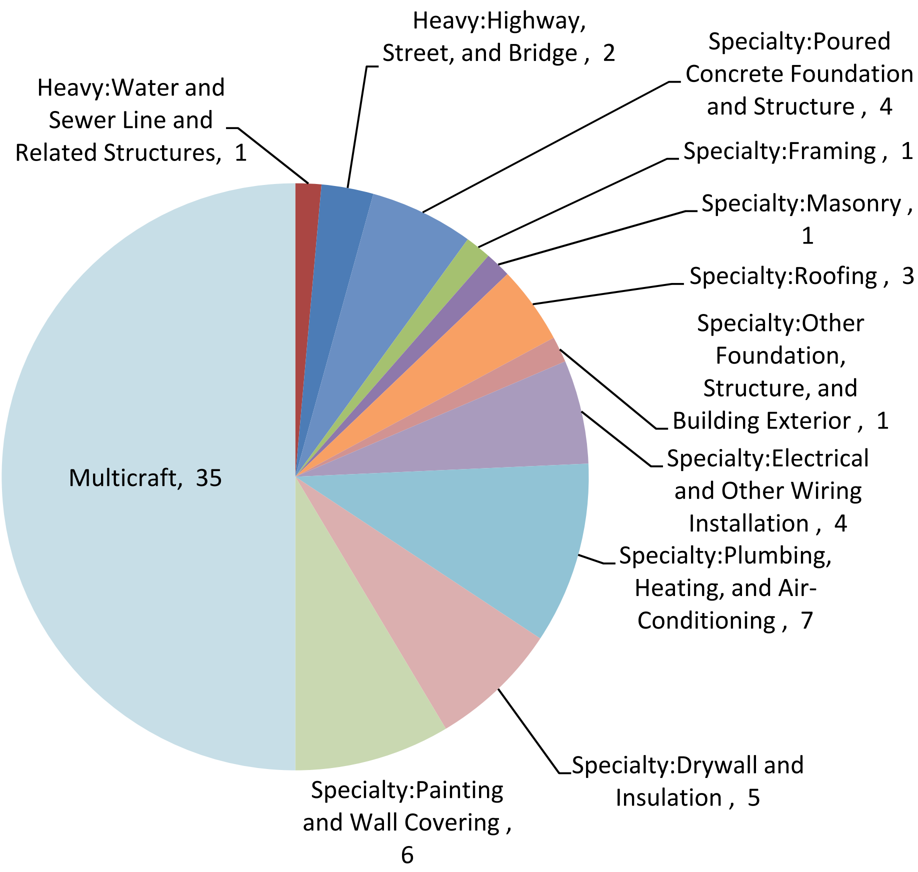

Figure 2 The ratio of violations by work type generalizes and scales from the San Jose metropolitan region to the entire state

Using construction as a case study: I could map 40% of the DOL cases from NAICS to SOC format, the other 60% map from 1 NAICS code to many undefinable SOC codes. For example, NAICS:Residential_Construction maps to 45 SOC codes such as electrician, roofer, framer, mason, etc. 30% of the DOL dataset is essentially coded as 'Residential & Commercial Construction. In the best case both codes should be used by the DOL so they have a code looking like 23.6117-47.4031 for Building:New Housing Fence Erectors. You could find these workers by visiting a new housing subdivision development and look for the guys building the fences. There is a good chance these guys are paid in cash, are not paid overtime, and are not having contributions made to their benefits. Knowing just 23.6117 or 47.4031is not enough to pinpoint fence erectors on subdivisions. Knowing just 23.6117 only tells me new housing and this can mean anything from fence erectors to roofers and electricians. Knowing fence erectors is more helpful since they cross multiple domains of NAICS codes.

One thing I found - in construction, half the cases are basically NAICS coded to 'other.' For example they have residential construction but this category of worker does not exist. The electrician on a residential project (NAICS 236) is just as likely to drive next to install some specialty equipment (NAICS 238) maybe on the same project where they would be doing NAICS 236 work. It makes it a hack assigning SOC employee population.

I suspect the residential/commercial building and the various 'other' categories are a catchall for 'multicraft' workers; they are electricians today, plumbers tomorrow, and do the roof at the end of the project. This is fairly common and difficult to then pinpoint the craft; I created a new category for 'multi-craft.'.

Mismatch in comparison between current BLS population and DOL violations over 20 years.

Average DOL data over 20 years for comparison ratios with current population. For example, the violations affecting pointers, take the mean number of painters affected by cases each year and compare this to the current population of painters, 100 painters over 20 years is 5 painters per year; 5/10,000 painters is the ratio we use[1].

Each row in the main whd_whisard dataset on the DOL data enforcement website is defined as "all concluded WHD compliance actions since FY 2007”. I have the following questions:

- I (Ash) see that the gap between “findings start” and “findings end” date range from 1985 to 2015 (there are some errors in date entry). My concerns are with matching up the numbers by NAICS code across years. An industry could have violations that went as far back as the 80s, but if we compare, for example, “EEs employed in violation” with external datasets from the SUSB or CBP (e.g. “total employees in industry group”) then isn’t it somewhat misleading? Say we’re using 2014 industry aggregate numbers by NAICS code. Then something like “Backwages per employee” when compared to aggregate 2015 numbers doesn’t seem right because these back wages were owed over a 25-30 year span. We’re essentially comparing past violations to current numbers to create any ratios, and this could be misleading in subsequent analyses. Does this make sense? I’d love to hear your thoughts.
- Assuming this is a cause for concern, a possible solution could be to take the average of industry statistics over the same 25 or so year span, and then calculate any ratios against industry numbers. This sort of analysis, in essence, will discount any variations in an industry group’s violation patterns over time.
- FP: For California the first cases are 2002 and the last are end of 2015; there are a few earlier cases but these look like investgations that started in 2002 or 2003 with an earlier violation start date; assumed based on number of cases in 2002 (111) versus 2001 (37), each preceding year has vastly fewer cases. This is 15 years.
- Averaging the violations by 15 years seems too much. In the San Francisco-Redwood City-South San Francisco, CA metro region there are 425 construction (div 23) cases. Currently there are 30,000 construction workers in this region. Just due to retirements we can assume half those 30,000 workers were in construction 15 years ago, the other half retired. Construction has a higher velocity of turnover entering and exiting the industry (citation, xxxx) due to the rigorous conditions. There for we can assume 15 years is excessive. My intuition is six to twelve years is the mean construction career (citation, xxxx). Of the 30,000 current workers, I would assume three years ago the majority were the same 30,000. Six years ago, man are but before six years ago, thirty percent are probably new workers.
- I propose we take the mean citations as over six years. For example, 425/6 = 70 and then 70/30,000 = 0.2% of workers in this metropolitan area have been part of a DOL case. 425/15 =

[1] Santa Clara wage Theft Coalition recommends a 2year average

28 violations and so 28/30,000 0.1% violations per worker seems too much of a discount. I agree that 425/30,000 = 1% of workers have been part of a DOL case is unrealistic given this is over 15 years. A six year interval seems fair.

- Ash thought comparing 15 years of cases to current population is a mismatch; I think he is using the mean cases per year, I am using mean cases for 6 years. So, mean cases per 6 years normalized by # of workers in that region. We are ignoring population growth. The idea is the construction worker standing here today was not the same dude standing there 15 years ago so a DOL case from 10 years ago means squat to his working conditions
- My six year intuition is based on: From my LiUNA survey, the average Union laborer has 10 years. 1/3 are under five years and 2/3 are over 5 years. The Laborers are thought to have the highest velocity in turn over simply because it is difficult. Non-Union usually have a higher velocity in turnover, if the Laborers are the benchmark then non-union are probably on average less than 10 years in a trade before they move on. The foreman and supervisors have more longevity (the Judas fucks they are) but they know the rest will turn over every ten years. I think I have seen supervisors that assume they can get away with it because they know you won't be around in another five years. If there was a recent action, a few years, they shape up and treat us better.

Checks

DOL case where these is no BLS population – none found in Bay Area Metro regions

Construction Bay Area Dive: Heatmap by Region and Occupation

Mean_wage = average of mean wage and median wage

DOL_ratio = ratio of DOL_annual_case to regional occupation population

DOL_annual_case = DOL cases total divided by 6 (years)

Heat Index = b ((100-mean_wage)/100) + b (regional_ee/10,000) + b (1-(DOL_ratio))

To clean up the duplicate NAICS to SOC codes I went through and by hand looked for NAICS codes violations that have no SOC coded population – indicating that the NAICS is duplicate (can't have a violation without workers).

Missed data

- Salinas BLS dataset is missing but has DOL cases
- Napa has BLS but no DOL cases

Conclusion

In my construction dive (using myself as an expert review for the reality check): I found half the violations/complaints involved workers categorized as no specific category. I also found occupations that had no employees in a metropolitan region - Ash found similar gaps. I think the gaps might be filled by undocumented workers. For example there are no bricklayers in one of the metro regions, this is illogical (where my expert review comes in), I suspect all the bricklayers are undocumented in that region. I ran these two ideas past Ash and he agreed they are something to look further into. The conclusion is we will probably find similar patterns of workers categorized as no specific occupation and gaps in employment in other industries. I don't have an expertise in those other industries so I would not be able to spot these inconsistencies as easily though if we agree these are valid observation we can assume they are there in the other industries.

Meetings

7/29/2016 Technical

- Ash
- Michael (RMBX)
- Chris P.
- Arturo (FFC)
- Forest

Notes

- No NAICS codes, no occupation employees in metro region
- 50% of complains have no category representation
- Source is survey metro, regional, state level
- Looking for groups that are not in the WHD dataset
- Exploitation severity score (# of complaints as a percentage of total workforce) * Backwages owed * size of workforce
- County Business Patterns (census bureau) (CBP)

To do

- Look for workforce demographic by NAISC
- Need demographic data for predictive models

7/22/2016 Technical

canceled

7/11/2016 Strategy

Monday at 5:30 p.m. on July 11 at the Law Center 1030 The Alameda, San Jose, CA 95126
This meeting is part of the regular Wage Theft coalition Meeting

- Forest
- Michael Tyag
- Ruth
- Wage Theft coalition members

6/22/2016 Technical

2:30PM San Francisco, The Iron Cactus on 4th street

The agenda is still forming; Ash has worked on an ROI calculator and is digging into the data structure; I took a dive into South Santa Clara County construction data. I think we are getting to good questions about the data and an understanding of the context and separations between the DOL and California Labor Commission datasets.

This meeting will be a technical data meeting - we are coming up on having a 'monthly' type meeting to present and get feedback on direction.

- Forest

- Michael.DOL
- Thor
- Ash
- Nathan
- Matt

6/19/2016 Informal

Informal meeting at Stanford to review to date progress and direction forward

- Forest (CIFE)
- Ting-Po (Statistics for Social Good)

6/16/2016 Informal

Informal conference call meeting

- Ash
- Michael (DOL)

6/6/2016 DataKind Project Accelerator

- Michael M. (RMBX)
- Michael (DOL)
- Matt (DOL)
- Forest (CIFE)

6/2/2016 Technical

11am at Stanford Y2E2 Coupa Café

Attending

- Kris (Statistics for Social Good)
- Ting-Po (Statistics for Social Good)
- Forest (CIFE)
- Michael (DOL)

General notes

We prepared for the DataKind event by discussing the Big Problem, the current data available, what has been done with the data, and key questions

5/18/2016 Initial Technical

11am at Stanford Y2E2 Coupa Café

Attending (*on email chain)

- Kris (Statistics for Social Good)
- Forest (CIFE)
- Marc (CIFE)
- Martin (CIFE)
- *Michael (San Jose City)
- Michael (RMBX)

- *Ruth (WTC)
- Michael (DOL)
- *Susana (DOL)
- *Jesse(FFC)

General notes

DataKind

- DataDive Hackathon: End of June, two day event that could produce visualizations; requires Michael/Forest to attend for two days – no problem. Kris will propose to DataKind. The DataDives are like "hackathons" for social good; they bring together data science volunteers from around the Bay Area for a weekend to investigate questions / prototype tools that DataKind leaders and organization partners identify in advance.
- Project Accelerator Night: Lower investment -- it's more like a group brainstorming session between local data scientists and partner nonprofits. DataKind has an Accelerator Night coming up soon (tentatively June 6, 6:30pm, in SF).

5/25/2016 Proposed FFC & WTC

The FFC is looking forward to meeting the WTC

Partners

Stanford

CIFE

Website http://cife.stanford.edu/

Contact

Forest Peterson granite@stanford.edu
Prof. Martin Fischer fischer@stanford.edu
Marc Ramsey mramsey@stanford.edu

Background

Forest Peterson is a PhD Candidate at the Center for Integrated Facility Engineering (CIFE) at Stanford. He is very interested in workers' rights issues since his mother is a workers' comp attorney who helps injured workers. I explained to him that the Department of Labor has data online on wage theft cases, but it isn't in a form that it easily accessible to the public by city or by industry (restaurant, care home etc.) Forest indicated to me that he was interested in working on this project. He was wondering if there was any money so he could also enlist a Masters' student to help, but, if there was not, he was willing to work on it himself. I thought that Forest, who has agreed to work on the DOL project, might want some help from Kris' and his volunteers at Stanford's "statistics for social good" so he wouldn't have to do it all himself.

Statistics for Social Good

Website http://stanford.edu/group/stats-for-good/

Contact

Kris
(805) 428-7231
sankaran.kris@gmail.com>

Data

I can see how including data can be used to supplement advocacy.

- Linking to industry information might be tricky
- Ruth’s other ideas seem doable
- Highest priority metrics for Statistics for Social Good related to these data sets
    - Scope
    - Specific analysis goals
    - Questions
    - Tasks
    - Types of data analysis
    - Organizing
    - Sharing
    - Statistics for Social Good Considerations

- Impactful to the Wage Theft Coalition / the Law Center / workers
- Whether Statistics for Social Good can meaningfully contribute

Background

I would like to e-introduce you to Kris Sankaran who is a Ph.D. candidate at Stanford who works with a couple of volunteer groups:"statistics for social good" a Stanford group and also DataKind.

The Santa Clara County Wage Theft Coalition has asked Kris to help us organize and make accessible the state Department of Labor Standards Enforcement data which I sent to him on a spreadsheet, but I think it would be good if he also had the DOL data. Thanks so much Kris for your willingness to meet with Michael and identify potential projects for Statistics for Social Good.

I (Kris) want to clarify that I am most likely not going to be contributing much to this project personally. My role in Statistics for Social Good right now is to identify potential projects, scope specific analysis goals, and recruit / prepare students to contribute to them. That said, the more details I can learn about what types of data analysis, organizing, and sharing would be impactful to the Wage Theft Coalition / the Law Center / workers, the easier it will be to attract data volunteers and set them up for a successful project.

DataKind

lilyhzhang1029@gmail.com
quale@cisco.com
majacaci00@gmail.com
kathleentang28@gmail.com
Ash (ashirwad08@gmail.com)
Nate (natemiller77@gmail.com)
Brian (brian.spiering@galvanizo.com)
Michael Myers, Data Scientist (michael@rhumbix.com)
Matt Giguere, Data Scientist (matt@rhumbix.com)

Wage Theft Coalition

Website https://wagetheftcoalition.com/

Contact

Ruth Silver Taube
Supervising Attorney
Workers' Rights Clinic
Katharine & George Alexander Community Law Center
1030 The Alameda, San Jose, CA 95126
Santa Clara University School of Law
(408) 737-2313
rsilvertaube@scu.edu

Database (California Labor Commissioner)

This dataset was obtained through a public information request and relates to the State of California. The California dataset is from the Labor Commissioner's Division of Labor Standards Enforcement (DLSE).

The division between DOL and DLSE cases is not simple – there are different criteria for coverage that the DOL uses and can have cases with smaller businesses. On the DLSE side, they can handle larger cases.

A course summary:

The Dept of Labor dataset is restricted to violations where the employer has more the $500k in revenue. The DLSE has cases where the employer has less than $500k in revenue. It is up to the complaint to file with the DOL or the DLSE, or both. There are differences in the restitution available through the DOL and the DLSE so that affects the decision process.

It seems like the DOL dataset will appear to have 'gaps' in the data but it is simply the result of another organization (like the DLSE) handling that subdomain of violations. I think we will need to be vigilant to fill these gaps with the relevant datasets. The foundation for Fair Contracting (construction specific) and the Santa Clara County Building Trades Council (construction specific) have datasets relating to violations they have handled. I assume these cross over both the DOL and DLSE datasets so there will be duplicate entries. I suspect they have entries in their database that for various reasons are not in the DOL or DLSE data. We don't have the FFC or BTC data yet so it is not something immediate to worry about.

Likely each industry subdomain will these regional associations that have subsets that will fill gaps in the hierarchy of datasets. If we can overlay these and look we will then have a mosaic of data that should form a complete dataset. The goal is to then look for gaps where employers could be flying under the radar. Out of those gaps we then need to look at the subset where we find high levels of the factors for exploitation.

There is no dollar amount floor for the State of California as you state. The $500,000 is just a federal floor. The data from the DLSE is for all companies. I use the DLSE process almost exclusively because there are waiting time penalties, more generous overtime, and meal and lunch break premium pay (available) under state law; we do file at the DOL in some cases since there is individual liability and, until this year, there was no individual liability in California. The DLSE information is from the Wage Adjudication Unit. There are hearings (Berman hearings) before a hearing officer, and then an ODA (order, decision, and award) is issued. If the employer does not appeal or pay, then the Labor Commission (DLSE) records the judgment in Superior Court in Santa Clara County. The judgements are the dataset you have.

The DLSE also has a Bureau of Field Enforcement (BOFE) that issues citations. The data does not capture the citations. I could request them, but right now we are focused on the judgments because we think the public can understand the necessity of paying judgments entered in court. The DOL issues final administrative decisions, and, since that's the last step in their process, these final administrative decisions have the same force and effect as the judgments when it comes to the wage theft policy and Ordinance that was enacted in the City of San Jose.

We are particularly interested in obtaining data on restaurants since we are advocating for health permit revocation of restaurants in the County for unpaid judgments. We are also interested in pinpointing the city where the violations occur and the number of violations in the city since we plan to go to other cities in the South Bay and advocate for wage theft policies, and these cities are interested in local data (from the cities) as well as industries.

For healthcare breakdown between 6-bed and 12-bed

Some datasets do not have the occupation, the application form the complaint completes usually has a field for occupation but is not entered in the system; this requires scanning the original forms to extract the occupation field

My phone no. is 408-737-2313 (cell: 408-621-5678)

Analysis and Visualization Goals

Michael Tayag, please chime in, but my opinion is that the Wage Theft Coalition/Law Center/workers with whom we work are most interested in

- Wage and hour violations by city and by industry (restaurant, care home, construction, etc.)
- Our focus currently is not on the other laws that the Department of Labor enforces.
- We are interested in state agency (Department of Labor Standards Enforcement (also called DIR (Department of Industrial Relations) and Labor Commission) as well as federal DOL (Department of Labor) data in the South Bay.

The Coalition is interested in industry data and data by city because we plan to go to cities to try to get them to enact wage theft policies and ordinances. The restaurant industry is particularly key at this moment because we are focusing on revocation of health permits at the County level for unpaid judgments. The construction industry, care home industry, and other industries are also important.

The DLSE data consists of all wage theft court judgments the DLSE recorded in Santa Clara Superior Court after hearings in which the hearing officer found in favor of the worker. There are often larger claims because the labor code in California provides for overtime (1.5 times regular pay) after 8 hours a day and 40 hours in a week and double time after 12 hours in a day. At the DOL, the worker can only recover for overtime after 40 hours in a week. In addition, the federal minimum wage the DOL enforces is $7.50 per hour. The DLSE enforces the local minimum wage of $10.30 per hour in San Jose ($11.00 per hour in Santa Clara) or the state minimum wage of $10.00 per hour where there is no higher local minimum wage. The DLSE has a wage adjudication unit and a Bureau of Field Enforcement. The wage adjudication unit holds hearings, and, as mentioned above, issues an Order, Decision, and Award (ODA). If the ODA is not appealed, the DLSE enters a judgment in Superior Court. The Bureau of Field Enforcement (BOFE) investigates cases involving more than one employee and issues citations. The data I provided through my Public Records Request does not include BOFE citations. (I may do a Public Records Act request to obtain that data). We are interested in both the DOL and DLSE data. It would be good if we could separate the data, but, if you wish, you can also aggregate them in addition.

Background

I formed the Santa Clara County Wage Theft Coalition because of my frustration at our low income, primarily immigrant clients who had DLSE wage theft judgments that were never satisfied. We have had rallies in front of the worst violators and have been advocating at the local government level for permit and license revocation as well as disclosure of unpaid judgments and nonissuance or revocation of contracts.

I wrote this op ed that that appeared in the Mercury News yesterday (5/16): http://www.mercurynews.com/opinion/ci_29899518/taube-bitmicro-latest-silicon-valley-firm-caught-victimizing

We plan to go to the various cities and get them to enact wage theft prevention policies and ordinances. We need to show them the magnitude of the problem with statistics and the industries that are impacted. It is easier to get legislators to enact legislation if there are judgments or final administrative decisions.

The WTC & FFC work overlaps.

This project will be extremely useful to the Coalition, the Law Center's clients, and workers.

Department of Labor

Website https://www.dol.gov/whd/

Contact

Michael Eastwood
Assistant District Director
U.S. Department of Labor
Wage and Hour Division
96 N. 3rd St. Suite 400
San Jose, CA 95112
(408) 282-4761
Eastwood.Michael@dol.gov

Blanco, Susana - WHD <Blanco.Susana@dol.gov>

Database

The Department of Labor has data online on wage theft cases, but it isn't in a form that it easily accessible to the public by city or by industry (restaurant, care home etc.) Organizing the DOL data and making it accessible by city and industry etc. will be a big (unofficial) project.

The DOL raw data on website: http://ogesdw.dol.gov/views/data_catalogs.php The wage and hour compliance data (at the bottom) has information on violations of federal minimum wage, overtime, family medical leave act violations, government contract violations, H-1B, H-2B, and H-2A violations, and more. That is the dataset that I have specific expertise in although the other datasets might also be quite interesting to you. You can also search the data online and see what DOL has already done to make the data accessible: http://ogesdw.dol.gov/views/searchChooser.php It’s my feeling that much more could be done and since the underlying data is public, anyone can do it.

See USB data.

SanFranciscoDistrictDOLCasesAndEmploymentCensus2016-07-06 - Underlying WHD Enforcement Data

- Each row in the main whd_whisard dataset on the DOL data enforcement website is defined as "all concluded WHD compliance actions since FY 2007". I have the following questions:
  - Q: I see that the gap between "findings start" and "findings end" date range from 1985 to 2015 (there are some errors in date entry). My concerns are with matching up the numbers by NAICS code across years. An industry could have violations that went as far back as the 80s, but if we compare, for example, "EEs employed in violation" with external datasets from the SUSB or CBP (e.g. "total employees in industry group") then isn't it somewhat misleading? Say we're using 2014 industry aggregate numbers by NAICS code. Then something like "Backwages per employee" when compared to aggregate 2015 numbers doesn't seem right because these back wages were owed over a 25-30 year span. We're essentially comparing past violations to current numbers to create any ratios, and this could be misleading in subsequent analyses. Does this make sense? I'd love to hear your thoughts.
  - Q: Assuming this is a cause for concern, a possible solution could be to take the average of industry statistics over the same 25 or so year span, and then calculate any ratios against industry numbers. This sort of analysis, in essence, will discount any variations in an industry group's violation patterns over time.
  - A: As far as the start date and end date, while there will be a few older cases that concluded in 2007 or later that were from much earlier times (due to litigation, etc.), the vast majority will be in the more recent time period of 2007 to the present. I would suggest limiting analysis to cases with findings end date of between 2008 and 2015. That more contemporary time period will match up better with census data. I wouldn't be too concerned about the difference of 7 years. If we have seen a pattern of violations in a given industry through our prior cases, then what we will want to know is how many EEs currently are in that industry for example. If you're really worried about that, you could also run the analysis limiting the data to 2011 to 2015 and see if that changes things in any significant way.
- For the SF jurisdiction specific dataset
  - Q: How are the "Minimum Wage Backwages per Employee" (and"Backwages per Employee") numbers are calculated? As an instance, if I test Minimum Wages Backwages / EEs Employed in Violation (or by Employees in Industry) I don't get the value in the "Minimum Wage Backwages per Employee" column, at least for the first row of the spreadsheet.
  - A: As far as how the "Minimum Wage Backwages per Employee" column was computed, I went back and checked and the way I did it really isn't right. I computed the average Minimum Wage Backwages per Employee on an individual case level. Then as part of a later SQL query to aggregate the data by industry I took a straight average of that column which means I really should have titled the column "Minimum Wage Backwages per Employee per Case" or something like that. The problem of course is that each case is given equal weight, regardless of the number of employees, so one case with 1 employee who had $0 in minimum wage backwages would have the same weight in my

> average as another case with 1000 employees each due $1000 in minimum wage. My "Minimum Wage Backwages per Employee" computation was $500, the average of the two cases, when it really should have been almost $1,000. I should have done a weighted average or, even more simply, just did what you did which is compute the average at the end based on total number of employees in the industry due back wages and the total minimum wage amount. I did this correction in the attached updated file and also included the raw underlying enforcement data. The underlying data is just a subset of the WHD Enforcement database.

Background

The federal Department of Labor handles wage and hour violations involving violations of federal wage and hour law (and a few other laws); the California Department of Labor Standards Enforcement (Labor Commission) is the state agency that handles violations of state wage and hour law. Both are very critical.

Unfortunately, we do not have grant funding available but we do have a ton of very interesting publicly available data that up to this point has gone almost completely unused. I can walk you through the data, explain some of what it means, and show you some examples of different ways in which the data could be used.

DOL Laws https://www.dol.gov/general/aboutdol/majorlaws

Foundation for Fair Contracting (FFC)

Website http://www.ffccalifornia.com/

Contact

Bryan Berthiaume
Executive Director
Foundation for Fair Contracting
3807 Dasadena Avenue, suite 150
Sacramento, California 95821
(916) 549-6380 cell; (916) 487-7871 office; (916) 487-0310 fax

Arturo Sainz
Bay Area Representative
916-549-6378 cell

Database

The database comprises cases the FFC has investigated over the years. This includes cases we have forwarded on to DIR, agencies, or lawsuits to ensure compliance.

The Wage Theft Coalition is interested in FFC cases with judgments (the DLSE recorded the judgment in Superior Court) or final administrative decisions (DOL, Office of Equality Assurance) in the San Jose offices of these agencies or even lawsuits that went to verdict or with settlements.

The Wage Theft Coalition is also interested in the kind of cases the FFC is seeing. Are they just wage and hour cases or are there cases that you filed at the EEOC/DFEH/NLRB.

Background

The FFC is one of the oldest fair contracting prevailing wage monitoring programs in the US.

Relevant Organizations

Santa Clara District Attorney Premium Fraud Unit

Website *

The FFC has recently formed an alliance with the Santa Clara District Attorney Premium Fraud Unit to help combat wage theft including premium fraud in particular.

I am glad the FFC has an alliance with the Premium Fraud Unit. The WTC has been very frustrated with this unit because they don't have a particular interest on wage theft unless there is workers' comp fraud (which there typically always is), but they aren't particularly interested.

California Department of Labor Standards Enforcement

Website http://www.dir.ca.gov/dlse/

Also called

- Labor Commission
- Department of Industrial Relations (DIR)

California Department of Labor Standards Enforcement (Labor Commission) is the state agency that handles violations of state wage and hour law.

City of San Jose

Website https://www.sanjoseca.gov/index.aspx?NID=144
Michael Tayag mtayag3@gmail.com

Santa Clara Building Trades Council (BTC)

Website http://www.scbtc.org/
Josue Garcia (CEO) offered his BTC dataset for our use.

Relevant Literature

News

- http://www.mercurynews.com/business/ci_29899028/tesla-responds-this-newspapers-investigation-imported-workers-at
- http://www.sfgate.com/news/article/Wage-theft-a-scourge-for-low-income-workers-2354262.php