Investigating Possible Locations of Undisclosed Subcontractors through Data Analytics of Employment and Federal Contract Records

M. Eduard Tudoreanu
University of Arkansas at Little Rock, 2801 S University Ave, Little Rock, AR
Phone 501-683-7268, Fax 501-683-7049
metudoreanu@ualr.edu

Keith Franklin
University of Arkansas at Little Rock, 2801 S University Ave, Little Rock, AR
Fax 501-683-7049
kxfranklin@ualr.edu

Ningning Wu
University of Arkansas at Little Rock, 2801 S University Ave, Little Rock, AR
Phone 501-683- 7051, Fax 501-683-7049
nxwu@ualr.edu

Richard Wang
Massachussetts Institute of Technology & University of Arkansas at Little Rock, 2801 S University Ave, Little Rock, AR
Phone 617-304- 3120
rwang@mit.edu

Abstract

This paper analyzes data from employment and Federal contracts, and it provides a characterization of how contracts affect local employment. It is a continuation of the research presented at the last year's ARP Symposium on searching for undisclosed subcontractors by exploiting the linkages between publicly available employment data and cuts to federal contracts. Many Federal and DoD contracts are performed by a team of contracting entities, where some prime contractors rely on subcontractors to execute specific parts of the contract. For many reasons, including national security, privacy, or competitive advantage, some of these subcontractors are not publicly disclosed and have the potential to be unmitigated single stress points in the acquisition process. The paper focuses on gaining a deeper and data-verified understanding of the interactions between federal awards and employment numbers, particularly on the boost to local employment that the start of a large contract may provide. A process for analyzing large Federal contracts side by side to employment information is presented. The result of the analysis has found that locations of large Navy awards rank above 70% of other locations in the country in terms of the magnitude of employment changes, under certain industry classification reporting methods.

Keywords

data analytics, purchasing data, FPDS NextGen, subcontractor, employment data, location quotient

Introduction

The ecosystem of DoD contracts often involves multiple entities, both large enterprises and small businesses, who work together to achieve the results needed by the DoD. Federal contracts can be awarded to single entities, but also to multiple contractors. Furthermore, contractors rely on other entities, henceforth subcontractors, to perform specific parts of the contract. For reasons of security, confidentiality, or competitive advantage, some of these subcontractors are not publicly disclosed. One overarching research question is whether enough publicly-available data exists to allow for the discovery of undisclosed subcontractors.

The open society and transparent government of the USA, though contributions from local and federal government as well as private companies, makes information available on a wide variety of topics, from air quality to social interactions, from Federal contracts to employment status. It is unlikely that a single such data repository would allow information about hidden subcontractors to be determined, though the combination of multiple data sources might do just that, especially if the same undisclosed contractor is participating in multiple contracts.

The data-driven approach taken in this paper is to analyze a large number of contract event, specifically the start, end, and any modifications. The reasoning is that one such event, for example securing a contract to perform work for the federal government may lead, at least in some cases, to a boost in employment in those locations closely related to the performance of the work. Similarly, the ending of an award or a negative modification may result in a drop in employment. When a large number of events are processed together, it may become possible to hone in the location of an undisclosed contractor as shown in Figure 1. This process could potentially be used to determine the likelihood of a location to be home to an undisclosed contractor for a given industry. A number of factors will influence the ability to detect the correlation between awards and employment, but two are worth considering next. First, the larger the award or award modification, the more likely it is to produce an effect on the employment data. Second, employment variations can be better detected in smaller cities and rural areas than in large metropolitan regions.

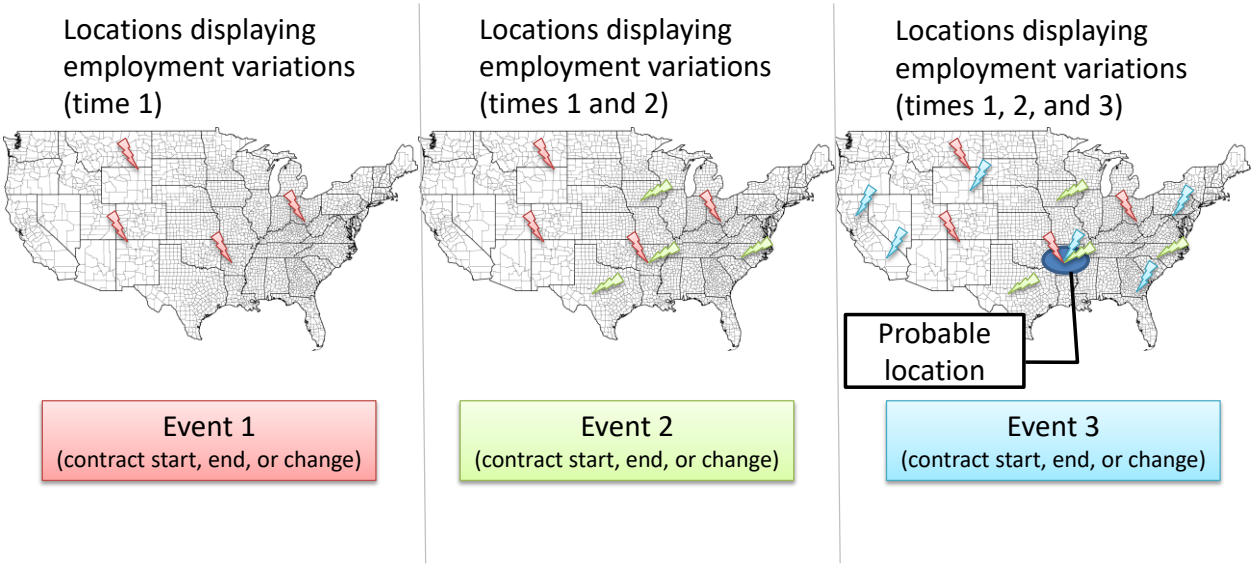

Figure 1: Determining possible locations for subcontractors based on repeated correlations between contract events and employment variations. Undisclosed contractors who participate in many awards, and thus pose more risk to acquisition, are more likely to be discovered.

The paper focuses on the narrower issue of determining the relationship between known contract events (beginning, end, modification) and employment variations at or around the same time both

a) at the disclosed locations from the contract, which provides a known relationship that can be reliably assessed, and
b) in the rest of the country, which will provide a measure of the existing noise in the data.

That noise can be further reduced by examining only industries related to each individual contract being studied and by considering employment trends at a local level in the context of the country as a whole.

The paper presents a process to analyze two large open data sources, one about federal contracts and the other about employment in the USA. Based on the datasets derived through the process, this research determined that a promising metric that relies on the magnitude of employment changes at a location and on a classification of industry type. On the datasets for the year 2016, the metric places locations that benefit from large contracts above 70% of other locations in terms of employment. Thus, through repeated elimination, it may be possible to narrow the location of an undisclosed contractor to a manageable number of places in the country, which can then be searched manually for web or social media presence of businesses capable of being an undisclosed contractor.

The rest of the paper covers related work in the next section, followed by an explanation of the methodology used to obtain and process the data. Results, conclusion, and future work are the topic of the last two sections.

Related Work

This paper expands our previous ARP Symposium work (Tudoreanu, Franklin, Rego, Wu, & Wang, 2018) that used a manual correlation process and revealed that reductions (modifications that cut the amount of money originally allotted) in contracts, which were large relative to the contractor's size, were correlated to a drop in employment in more than two thirds of the cases. Location quotient (LQ), which provides a relative measure of a region's employment in an industry sector relative to the nation, showed a correlation for 75% of the studied contracts' reduction. The main difference between this research and the past is the addition of positive contract modifications (including contract start date) into the analysis, the use of data science to be able to handle and process tens of thousands of Navy contracts, and the examination of the noise present in employment data.

Previously, policy makers and researchers have recognized the need to employ data as a multifaceted means of increasing the agility of the acquisition process (Krzysko & Barney, 2012). To this end research has looked at automatic means of dealing with the heterogeneous acquisition data sources from text processing (Zhao, MacKinnon, Douglas, Gallup, & Shelley, 2015), systems engineering (Cilli, Parnell, Cloutier, & Zigh, 2015), and business (Gaither, 2014) perspectives. Our paper is different both in content and in the approach. In content in that we seek to characterize hidden flow of funds in the supply network, and in the approach in that our expertise in data quality and data science provides a more value-based perspective.

A recent approach used the data collected by government agencies. The objective was to use administrative data i.e. state unemployment insurance information that is from covered wages and salaried workers based on the workers quarterly earnings. In studying domestic outsourcing, the primary constraint was how existing data is limited in estimating the number of workers who are employed by contractor companies or who provide services to firms as independent contractors. This data "can be used to document employment in contractor firms and the number of independent contractors, link contractor industries with the firms using their services." (Houseman & Bernhardt, 2017). The paper also recommends using earnings information that includes wages, salary employment, and self- employment tax data that can be provided by the IRS.

A second paper "uses linked worker-firm administrative tax data from U S tax returns to explore the changing relationship between firms and independent contractors." (Miller, Risch, & Wilking, 2017). A dataset was constructed, that used digitized tax filings from the IRS, for the tax years of 1997 to 2015. Individuals were linked to their employer via reports Form 1040, Schedule C and Schedule SE, and information reports W-2, 1099-MISC and 1099-K.

There have been other researchers who successfully integrated publicly available data with private information, but their goal was radically different than the scope of this paper. Such data integration efforts have been employed to reconcile corporate names (Gayo, Pablos, Rodriguez, & Vafopoulos, 2013) and to provide a tools for accessing unified corporate names (Llorens, Rodriguez, & Vafopoulos, 2015). Another approach involved government data

enhanced with information services from outside sources to provide additional, generalized context into the data (Felten, Robinson, Yu, & Zeller, 2009).

Methodology

Two data sources were used in our analysis, Federal Procurement Data Systems - Next Gen (FPDS) and Bureau of Labor Statistics (BLS). The procurement database provides a list of federal contracts awarded over the years, including any modifications to an award. The labor statistics includes employment data on a monthly and quarterly basis, both in absolute numbers and relative to the rest of the country, through the location quotient (LQ, 2008). Due to the limited amount of time and computational resources, the primary focus was limited to the year 2016, yet even this approach resulted in some operations involving tens of millions of entries and long processing times. The methodology described here is applicable to the analysis of additional years.

Data from the two sources was copied to a local database hosted on a relatively powerful server powered by an 8-core, dual Intel Xeon processor with 48 GB of memory. Java programs and MySQL queries were used to populate the database.

The data processing steps are listed below, along with an explanation of the types of data involved in those steps. The end-goal is three-fold:

a) to determine whether a large contract correlates with an increase in employment given the contract's industry and location,
b) to determine the overall behavior of employment in the entire US, and
c) to compare the employment at contract's location to all the other locations in the country at the same time (of the year) and industry in order to be able to determine a metric that has the potential to uncover hidden subcontractors.

To this end, steps 3 and 4 are split on two separate pathways, one to examine employment behavior strictly related to FPDS contracts, and one to study the employment trends in the entire country.

Step 1: Select contract data from FPDS

The analysis of contract data started with FPDS records for the Department of the Navy, and it underwent two selection steps. First, only contracts and modifications with a start or end date in 2016 were considered. Second, the analysis filtered "small" contract events (begin, end, or modification events) which in this case were events with dollar amounts in the range $-99,000 and $499,000. That is only changes to a contract that either reduced the amount awarded by more than $99,000 or had more than $499,000 were considered because they have a larger chance on producing layoffs or hiring. This step resulted in over 23,000 contracts (and modifications) being selected.

While a wealth of information is available for each contract, the relevant data to be used for the rest of the analysis includes:

- dollar amount;
- zip code of the principal place of performance, which can provide the local area in which to examine employment;
- start date, used as the effective date when the contract could start affecting employment;
- NAICS code, which is a standard way of classifying various types of industry in US and Canada (NAICS, 2017). The codes in FPDS are 6 digits long. NAICS uses a somewhat hierarchical structure where related sub-industry share the first few digits of the larger industry type. For example, a four-digit ABCD code is generally a sub-type of the three-digit ABC industry.

Step 2: Employment Data from Bureau of Labor Statistics (BLS)

Employment data included both monthly and quarterly metrics. Due to the nature of our search, the year 2016 was selected because it was already in our systems. The last quarter of 2015 and the first of 2017 were also considered for some of the steps below. The number of quarterly entries in the year 2016 alone is over 14 million (that computes to over 52 million monthly entries) for all covered locations and industry codes.

The following fields were deemed important:

- month of the year to determine the time frame;
- industry type, which was provided as NAICS (2017) codes. Three, four, five, and six digit industry codes were used (see Step 1 for an explanation of the NAICS code length);
- US county and state;
- number of people employed in the county per industry type and month. Note that not all counties have employment for all NAICS codes;

Step 3: Contract Pathway: Filter Relevant BLS Data Based on Large Contracts

From this point, two pathways were pursued in the processing. One that focused on employment relevant to an existing large contracts, henceforth termed Contract Pathway, and one that focus on the overall employment situation in all locations in the US, named All Locations Pathway.

This steps reduced the size of employment data (step 2) to only those locations that appear in at least one of the large contracts selected at step 1. The zip code of the place of performance of a contract was mapped to county and state. All monthly employment information about that location was copied in a working data store regardless of NAICS or time of year. All of 2016 and last quarter of 2015 and first of 2017 went through this transformation. For the year 2016, only about 2 million quarterly (that is 6 million monthly) entries were excluded.

Step 3: All Locations Pathway: Filter Relevant BLS Data Based on Large Contracts

While it is important to analyze the employment trend in the country as a whole, there is little reason to consider NAICS codes that are completely unrelated to any of the large contracts in the other pathway because employment performance in one part of the economy (say agriculture) is not directly correlated to performance in another type of industry (say information technology).

We designed an easily measurable definition for what would mean for two industries to be related. The FPDS data provides a six-digit NAICS code for the product that is the subject of the contract. BLS uses the NAICS code slightly differently in that BLS identifies the type of employer. It is possible that one employer may have products in different, yet related industry types. Our definition of related industries relies on the hierarchical nature of NAICS. Thus, in addition to the six-digit code known from FPDS, we also considered the, often more encompassing, five-digit code obtained by removing the last digit, as well as the four- and three-digit ones. Thus, for each FPDS contract, we created a bundle of related four related industries by the process of removing last digits of the code.

Given our definition of related industry types, this step removed all employment information for NIACS codes that do not belong in an industry bundle from at least one of the large contracts. Note that this automatically excluded codes with fewer than three or more than 6 digits. The result was a dataset of about 4.8 million quarterly (14.4 million monthly) employment entries. No data from years 2015 nor 2017 was included.

Step 4: Contract Pathway: Calculate Employment Before and After Contract Start Date

The start date of the contract may be a possible event that can lead to observable changes in employment levels. It is unlikely that any business would start hiring before the contract begins, and thus before funds become available. One good temporal point to be considered as the reference employment is the month right before the start date. We will refer to it as the *before* employment level. Similarly, after the contract is awarded, the business may hire additional people to perform the work. The hiring may actually take some time, so we decided to examine the contract start month and two more months after for any changes in the employment. Positive dollar amounts might lead to higher employment, therefore we use the maximum employment level reported in any of the three months (start month and the immediately following two months) as our *after* employment metric.

The before and after levels were computed for each contract and for each NAICS code from the related industry bundle (see step 3 All Location Pathways for an explanation of the bundle). An entry in this dataset consists of:

- contract data (dollar amount, location, start date);
- NAICS code (either a three-, four-, five-, or six-digit code). Not all combinations of location, code, and month have reported employment data;
- local before and after employment levels.

The dataset has over 53,000 entries, which include the multiple industry bundles, with about 16,400 unique contracts/modification. Some of the contracts from step 1 were not included because the employment information was not available for that location and time of the year either before the contract start date, or after, or both.

Step 4: All Locations Pathway: Calculate Employment Before and After Each Month

A similar technique was used to calculate employment levels in every county in the US for months starting in February and ending in October 2016. The before level is the one from the month before the month being calculated (thus the first one is February), and the after level is maximum of the three month following, and including, the target month (thus the last possible target month is October because it requires November and December data).

The dataset has the following fields, with more than 3 million entries. Note that it is possible that employment level was not reported for some locations, NAICS, and month combinations.

- month;
- NAICS code;
- county;
- before and after local employment levels.

Step 5: Combining the two pathways

This final step of data processing involves comparing each contract entry from step 4, Contract Pathway, to each possible location in US (step 4, All Locations Pathway). All possible combinations were generated as long as they had the same NIACS code and the same month. The total exceeds 52 million combinations. In additional to being able to perform a pair wise comparison, we also determined the national average before and after employment levels at the time of each contract.

Results

This section presents two broad metrics for employment levels. First, it just examines the whether the employment registered an increase as shown by the *before* level being smaller than the *after* numbers. Second, the magnitude of the change, whether increase or decrease, from *before* to *after* is presented. For this section, we focus on positive contract events, particularly a contracts start date and any positive modification to the dollar amount. The majority of the contracts, around 13,900, from step 4, Contracts Pathway, fall into this category.

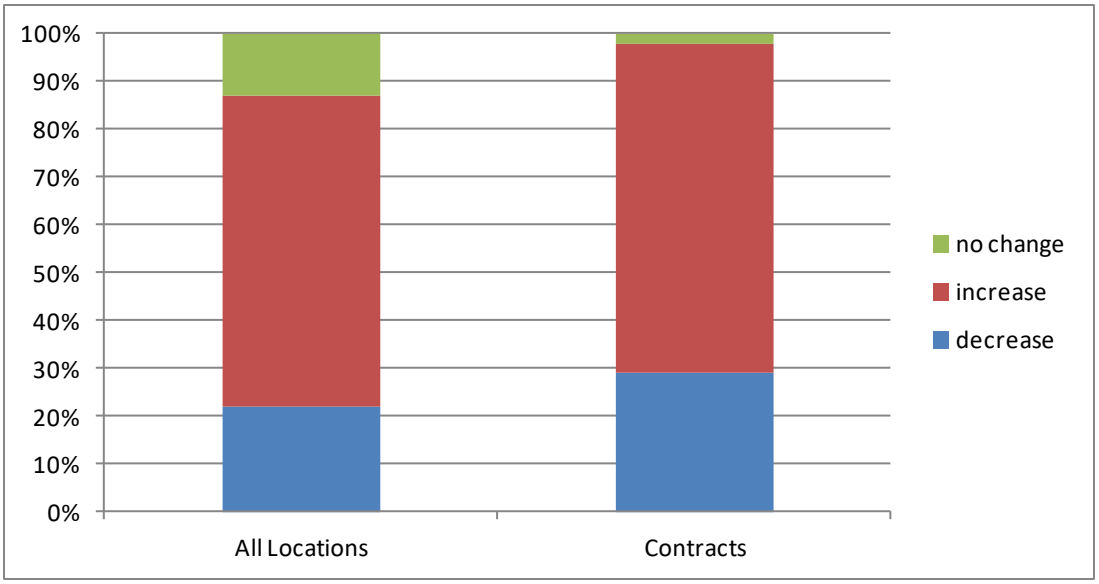

Figure 2: Relative changes in employment as recorded in the datasets produced in step 4. Both the country as whole over the year 2016 and those locations affected by awarded contracts show similar behavior in the number of instances where an increase is recorded (as a percentage of all entries in each dataset).

The country overall is experiencing an increase in employment during the studied period. Based on step 4, All Locations Pathway, in 65.2% of the counties and NAICS codes over the year, the employment records show an increase. For the locations, times, and NAICS codes related to contracts. that number is only slightly larger at 68.8% (from dataset produced at step 4, Contracts Pathway). The results are depicted in Figure 2.

A change in employment in itself does not seem to be sufficient to allow the discovery of an undisclosed contractor. Thus, a second metric, the magnitude of the employment change, was considered. Formally, the magnitude is obtained by subtracting the *before* number from the *after*. The magnitude can be negative if the employment level drops. Using the data from step 5, the magnitude in the counties related to a contract, at the start date of the contract, and for the relevant industry bundle was compared to magnitude of change for the same industries and at the same time for each and every county in the US.

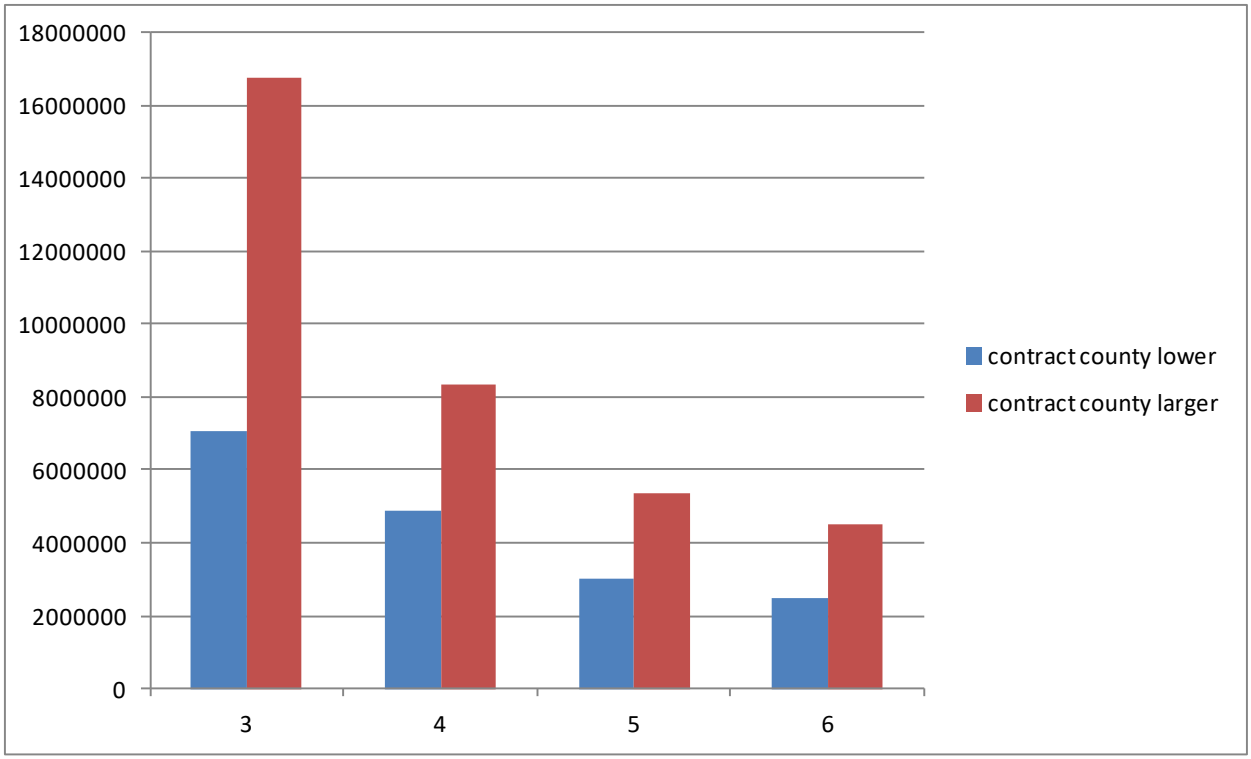

Figure 3: Comparing the magnitude of employment changes on contract locations, times, and NAICS code with all other recorded locations in the US. The bars' height shows the number of pairs in which the magnitude of employment change in the contract county is larger (red) or lower (blue) than some other county's. The x-axis breaks down the numbers by length of NAICS code.

Using the dataset from step 5, the locations of an awarded contract have a larger magnitude of employment increase when compared to all other US location (about 35 million out of the 52 million possible pairs). Figure 3 shows the result broken down by the length of the NAICS code, and it can be seen that for three-digit NAICS code, the relative percentage between larger contract locations (red) and lower ones (blue) is the highest at 70.3%.

Conclusion and Future Work

The paper employed large data analysis and found that the magnitude of employment changes is higher in 70% of places of performance for a Federal award as compared to the rest of the country for three-digit NAICS codes. This finding is promising as a metric to help uncover hidden contractors because it has the potential to eliminate more than two-thirds of US locations. It can be used repeatedly for many contracts that use products and services in related industries to assign a probability for various potential locations of undisclosed subcontractors. Such undisclosed contractors are typically employed in contracts with a larger scope to achieve confidentiality, security, or a competitive advantage. Depending on the situation, acquisition experts may need additional planning to protect such hidden contractors if security is desired, or may rely on data science to identify these contractors and avoid them becoming a weak link in the acquisition process.

The main contributions of this work are the development of a data science process for joining large acquisition and employment datasets and the development of a potential metric based on using the magnitude of employment changes in three-digit NAICS code industry. Future work will focus on three main thrusts:

a) running simulations where one contract is picked at a time, and the most likely locations of that contract are determined based on the three-digit NAICS code metric presented in this paper;
b) considering additional years, especially one where the employment is declining, which would be the opposite of 2016, a good year for employment; and
c) determining additional probability metrics for possible locations for contractors. Good candidates may be combinations of NAICS codes and location quotients (LQ, 2018).

Biographical Data

M. Eduard Tudoreanu is Professor of Information Science at University of Arkansas Little Rock. Professor Tudoreanu has expertise in human-computer interaction, information quality, advanced visualization of complex data, and virtual reality. He worked on visual data analysis, and has extensive experience in software development and user interface design. Professor Tudoreanu was the founding Technical Director of the Emerging Analytics Center. He has been the keynote speaker at ABSEL 2010, served as a panelist for the National Science Foundation and Missouri EPSCoR. He earned his Doctor of Science degree in Computer Science in 2002 from the Washington University in St. Louis.

Keith Franklin is a PhD candidate in Information Science at University of Arkansas Little Rock. He has over 20 years' experience in project management, IT, and Quality. He holds certifications in Project Manager (PMP), Information Security Manager (CISM), Risk and Information Systems Controls (CRISC), Quality Engineering (CQE), and Business Improvement (BIA). He has held management positions at Johnson & Johnson and SIMS Industries Medical, Pfizer (formerly Warner-Lambert division). He earned his Master of Business Administration from City University of Seattle and was honorable discharged from the U.S. Navy.

Ningning Wu is Professor of Information Science at the University of Arkansas at Little Rock. She received a B.S. and a M.S degree in Electrical Engineering from the University of Science and Technology of China and Ph.D. in Information Technology from George Mason University. Dr. Wu's research interests are data mining, network and information security, and information quality. She holds certificates of the IAIDQ Information Quality Certified Professional (IQCP) and the SANS GIAC Security Essentials Certified Professional.

Richard Wang is Director of the MIT Chief Data Officer and Information Quality Program. He is also the Executive Director of the Institute for Chief Data Officers (iCDO) and Professor at the University of Arkansas at Little Rock. From 2009-2011, Wang served as the Deputy Chief Data

Officer and Chief Data Quality Officer of the U.S. Army. He received hisPh.D. in information Technology from the MIT Sloan School of Management in 1985.

References

Cilli, M., Parnell, G. S., Cloutier, R., & Zigh, T. (2015) “A Systems Engineering Perspective on the Revised Defense Acquisition System.” Systems Engineering, vol. 18, no. 6, p. 584–603.

Felten, E. W., Robinson, D., Yu, H., & Zeller, W. P. (2009). Government Data and the Invisible Hand. Yale Journal of Law and Technology, 160

Gaither, Carl C. (2014) “Incorporating Market Based Decision Making Processes in Defense Acquisitions.” International Journal of Defense Acquisition Management, vol. 6, p. 38–50.

Gayo, J. E., & Pablos, P.O., & Rodriguez, J.M., & Vafopoulos, M. (2013). Towards a stepwise method for unifying and reconciling corporate names in public contracts metadata. The CORFU technique. 7th Metadata and Semantics Research Conference Track on Metadata and Semantics for Open Repositories, Research Information Systems and Data Infrastructures, Thessaloniki, Greece.

Houseman, S., & Bernhardt, A. (2017). Memorandum: Data Needs for Research on Domestic Outsourcing in the United States. Retrieved from https://research.upjohn.org/staff_publications/1.

Krzysko, M. & Baney B., (2012) "The Need for “Acquisition Visibility”." Journal of Software Technology 15-1, p. 4-9.

Llorens, L., & Rodriguez, J.M., & Vafopoulos, M. (2015). Enabling policy making processes by unifying and reconciling corporate names in public procurement data. The CORFU technique (Vol. 41, pp. 28-38). Madrid, Spain. Computer Standards & Interfaces.

Location Quotient, LQ (2008). *Bureau of Economic Analysis.* Retrieved from https://www.bea.gov/faq/index.cfm?faq_id=478.

Miller, A., Risch, M., & Wilking, E. (2017). Independent contractor or employee? The changing relationship between firms and their workforce and potential consequences for the U.S. income tax.

NAICS (2017). North American Industry Classification System. *US Office of Management and Budget.* Retrieved from https://www.census.gov/eos/www/naics/2017NAICS/2017_NAICS_Manual.pdf

Tudoreanu, M. E., Franklin, K., Rego, A., Wu, N., Wang, R. (2018). Searching Hidden Links: Inferring Undisclosed Subcontractors From Public Contract Records and Employment Data. In Proceedings of the 15th Annual Acquisition Research Symposium, Vol. 2, p. 491-499, Monterey, CA.

Zhao, D., MacKinnon, D., Douglas, J., Gallup, D., & Shelley, P. (2015). “Lexical Link Analysis (LLA) application: Improving web service to Defense Acquisition Visibility Environment” (NPS-AM-15-129). Monterey, CA: Naval Postgraduate School. September.