UALR-AM-20-156

ACQUISITION RESEARCH PROGRAM SPONSORED REPORT SERIES

Improve Acquisition and Procurement through Data Governance and Information Quality

September 14, 2020

Dr. Richard Wang

University of Arkansas at Little Rock

Disclaimer: This material is based upon work supported by the Naval Postgraduate School Acquisition Research Program under Grant No. HQ0034-18-1-0004. The views expressed in written materials or publications, and/or made by speakers, moderators, and presenters, do not necessarily reflect the official policies of the Naval Postgraduate School nor does mention of trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF DEFENSE MANAGEMENT
NAVAL POSTGRADUATE SCHOOL

The research presented in this report was supported by the Acquisition Research Program of the Graduate School of Defense Management at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact the Acquisition Research Program (ARP) via email, arp@nps.edu or at 831-656-3793.

Acquisition Research Program
Graduate School of Defense Management
Naval Postgraduate School

Abstract

The Federal Funding Accountability and Transparency Act of 2006 (FFATA) required federal contract, grant, loan, and other financial assistance awards of more than $25,000 be displayed on a publicly accessible and searchable website to give the American public access to information on what and how the federal government spends money every year. Federal acquisition databases, such as those maintained by USAspending.gov and FPDS.gov, serve this purpose. These databases contain contract information for all US Departments for the last twenty years. However, little has been done to examine the data and extract information that may provide valuable insights on potential ways to improve the efficiency of acquisition management. This report presents a data science approach to assessing and enhancing the quality of the databases and to discovering patterns that can be potentially useful for acquisition research and practice. Two key findings from data analytics include: 1) by utilizing publicly available weather information, we found that some zipcodes have high risks in natural disasters according to historical weather data since 1950, and some projects have a high percentage of contractors located in those high risk areas; 2) by clustering contractors based on their NAICS (North American Industry Classification System) codes, contractors are found to have a unique NAICS code with no other companies in the DoD's contractor pool capable of performing the same business or providing the same service. These findings provide ways in identifying risk factors in a project so that effective strategies can be designed to mitigate potential risks.

THIS PAGE LEFT INTENTIONALLY BLANK

UALR-AM-20-156

ACQUISITION RESEARCH PROGRAM SPONSORED REPORT SERIES

Improve Acquisition and Procurement through Data Governance and Information Quality

September 14, 2020

Dr. Richard Wang

University of Arkansas at Little Rock

Disclaimer: This material is based upon work supported by the Naval Postgraduate School Acquisition Research Program under Grant No. HQ0034-18-1-0004. The views expressed in written materials or publications, and/or made by speakers, moderators, and presenters, do not necessarily reflect the official policies of the Naval Postgraduate School nor does mention of trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

THIS PAGE LEFT INTENTIONALLY BLANK

Table of Contents

THIS PAGE LEFT INTENTIONALLY BLANK

Introduction

Defense acquisition consists of different data silos. These data silos have both technical and cultural origins. The capabilities to draw upon data across information systems hold huge potential for improving defense acquisition and procurement. Acquisition planning and management involves many decision-making and action-taking processes that cover a complex environment including actual acquisition, contracting, fiscal, legal, personnel, and regulatory requirements. A sound decision-making process has to rely on data – high quality data. Often the available data is dirty, outdated, incomplete, or insufficient for the expert to make a decision. On the other hand, there are enormous amounts of data on the Web that could be utilized to crystalize the needed information.

Our work investigated how to leverage information from public data sources to complement the internal data in order to support effective acquisition planning and management. The research is based on publicly accessible government acquisition databases from *usaspending.gov* and *FPDS.gov*. Both databases host federal spending data from the last two decades, and contain millions of records with detailed information about each contract. These rich repositories of data provide a great opportunity for us to learn from past practices, and to possibly gain some insights that can help with the design of better strategies for managing future projects.

A preliminary study (Wu, Tudoreanu, & Wang, 2018). showed that acquisition data suffer from quality problems, as do all other real-world data. Thus, to achieve accurate data analytics, the quality of data must be understood and improved. Our research demonstrated the feasibility of using online information from reputable sources to complete missing values and correct erroneous or inconsistent data in acquisition databases. The report takes a step further. It aims to enhance the acquisition data with online information so as to discover patterns that otherwise would not be discoverable.

Trust is a key issue for using online data. In fact, the Web has not only changed our ways of sharing and seeking information, it has also altered traditional notions of trust due to the fact that the information can be published anywhere by anyone for any

purpose, and there is no authority to certify the correctness of the information. It is often up to the information consumers to make their own judgement about the credibility and accuracy of information they encountered online. Unfortunately in the world nowadays, people are flooded with fake news and internet scam, thus it becomes even harder for an information seeker to discriminate between true and false information. To make the situation even worse, even when data are deemed trustworthy, assessing the data quality in big data era still faces many challenges. First, the diversity of data sources brings abundant data types and complex data structures, which increases the difficulty of data integration. Second, data change very fast and the timeliness of data is very short, which necessitates higher requirements for processing technology (Cai & Zhu, 2015).

This report only explores the usage of information from credible and reputable sources to enhance data analytics ability. However, investigating appropriate methods to assess Web data quality, to identify and acquire credible and accurate information is one of our ongoing research topics.

Two major findings that may help identify risk factors in an acquisition project are highlighted here. By utilizing the natural disaster data from the NCEI (formerly the National Climatic Data Center) and disaster assistance data from FEMA (Federal Emergency Management Agency), the project identifies areas that are prone to have more than one major type of natural disasters such as hurricane, tornado, flood, and wild fire. Contractors located in these areas are considered to have a high risk in natural disasters. It then studies the distribution of high risk contractors for each project and discovers that some projects have a high percentage of high risk contractors. These contractors could significantly impact the outcome of the entire project if they were struck by a natural disaster.

The type of business is identified by NAICS (North American Industry Classification System) code. A NAICS (2017) code can be attached to many products and many companies. However, if a NAICS code has only a few companies associated with it, then it can be considered as a high-risk business type, because if one of these companies failed, it would be difficult to find an alternative source. This information is

potentially important in the context of the NCEI and FEMA data to evaluate the severity of a disaster and its impact acquisition projects. Such a classification system is helpful to acquisition management for risk assessment. By clustering contractors by their NAICS, our research discovers that a number of contractors provide unique services/products to DoD projects and no other companies can replace them in the DoD contractor ecosystem. These contractors could be a possible weak link in a supply chain if they failed due to a natural disaster or malicious attack.

THIS PAGE LEFT INTENTIONALLY BLANK

Research Methodology

The research work follows the Data Enhancement and Analytics System framework shown in Figure 1 (Wu, Tudoreanu, & Wang, 2018). The system has four major components, namely Quality Assessment engine (QA), Data Cleaning (DC) Engine, Data Enhancement and Analytics engine (DAE), and Text Retrieval and Analysis engine (TRA). The key component is Text Retrieval and Analysis engine as it supports the functionalities of the other three components. TRA is responsible for four tasks: 1) performing searches on the Internet, 2) identifying the websites that contain the most reliable data, 3) extracting the desired information from the text; and 4) information fusion by collectively integrating information from multiple sources. When information for quality assessment and data cleaning is not available, TRA will search and extract the needed information online. Data Analytics and Enhancement engine aims to enhance knowledge about data by discovering hidden and interesting patterns in the data as well as complementing the internal data with the information that is not found in the database, but may be potentially useful for advanced data analytics.

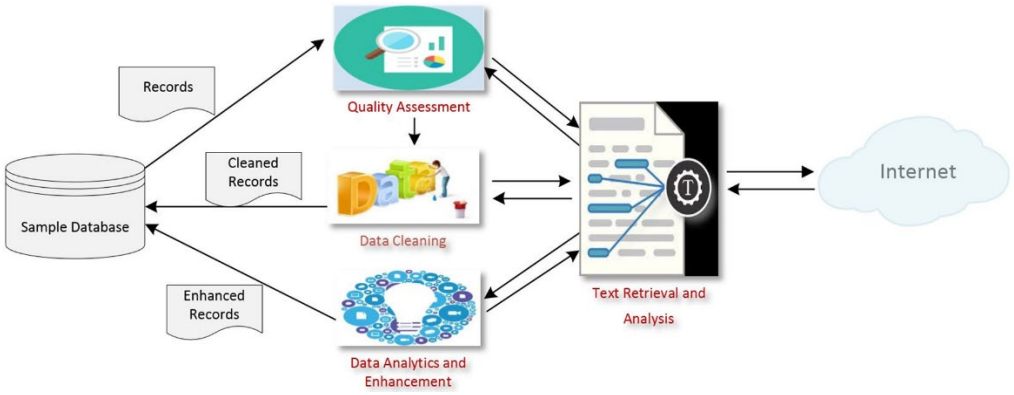


Figure 1: Framework of Data Enhancement and Analytics System

Our research methodology consists of the following two phases:

- Compare the data between FPDS.gov and usaspending.gov in terms of their structures, contents, and quality.
- Apply data analytics techniques to discover patterns from past acquisition projects. These patterns might help us to identify room for improvement in future projects.

THIS PAGE LEFT INTENTIONALLY BLANK

Comparison of FPDS and USASPENDING Data

Both usaspending.gov and FPDS.gov sites are publicly accessible and have contract information from all US Departments since 2000; however, the data in the two sites have different structures with different number of attributes. The data from usaspending.gov are categorized under prime award and sub-award. The types of spending include contracts, grants, loans, and other financial assistance. For each spending type, the data is organized into two structures: prime award and sub-award. For example, information on contracts is organized into two tables: one for prime contracts and the other for sub contracts. Data in FPDS.gov is available to download in an XML format, which has a hierarchical structure with various pieces of information such as award specifics, competition type, entities performing the work, and place of performance. For this analysis, the spending data from the Department of Defense were downloaded and stored on a MySQL database server.

Table 1 shows the structure of database tables from each source, where FPDS table is from FPDS.gov and all other tables are from usaspending.gov. Here, RecCnt and ColCnt represent the number of records and number of columns in a table respectively; CompleteCols and SingleValCols represent the number of columns with no missing values and number of columns with only a single value across all records; and EmptyCols and IncompleteCols represent the number of empty columns and the number of columns with missing values respectively.

Table 1: Profiling of FPDS and usaspending tables

Table Name	ColCnt	CompleteCols/ SingleValCols	EmtpyCols	IncompleteCols
PrimeContracts	221	50/1	0	162
SubContracts	101	41/0	3	57
PrimGrants	67	32/5	2	33
SubAGrants	101	29/4	25	47
FPDS	210	74/3	1	136

At a closer inspection, it appears that the FPDS table is more similar to PrimeContracts table from usaspending.gov in terms of their schema and contents.

Thus, the remaining part of this section compares only these two tables in terms of their schema, data coverage, and quality.

To facilitate the data comparison, attributes are classified into two categories: identity attributes and non-identity attributes. Identity attributes provide identity information for a contractor, contract, funding agency, etc. Examples of identity attributes include project identifier, funding agency identifier, contractor identifier such as DUNS number, business name, address information, phone, fax, etc. Non-identity attributes do not provide any identity information.

Schema mapping

Schema mapping between two tables is performed manually based on the data dictionary provided by each database. There are 180 common fields in two tables even though these fields are named differently in each table. The remaining 30 attributes in FPDS and 41 attributes in PrimeContracts are only found in their own table. Due to the space limit, Table 2 only shows partial mapping results.

Table 2: Schema Mapping between FPDS and primeContracts tables

	Mapping Attributes	
	Attributes in fpds	**Matched Attributes in PrimeContracts**
1	awardID_awardContractID_PIID	piid
2	awardID_awardContractID_agencyID	agencyid
3	awardID_awardContractID_modNumber	modnumber
4	awardID_awardContractID_transactionNumber	transactionnumber
5	awardID_referencedIDVID_PIID	idvpiid
6	awardID_referencedIDVID_agencyID	idvagencyid
7	awardID_referencedIDVID_modNumber	idvmodificationnumber
8	competition_A76Action	a76action
9	competition_commercialItemAcquisitionProcedures	commercialitemacquisitionprocedures
10	competition_commercialItemTestProgram	commercialitemtestprogram
11	competition_competitiveProcedures	competitiveprocedures
12	competition_evaluatedPreference	evaluatedpreference
13	competition_extentCompeted	extentcompeted
14	competition_fedBizOpps	fedbizopps
15	competition_idvNumberOfOffersReceived	numberofoffersreceived
	.....	.....
	.....	.....
165	vendor_vendorSiteDeta__ndorSocioEconomicIndicators_isIndianTribe	isindiantribe
166	vendor_vendorSiteDeta__allyDisadvantagedWomenOwnedSmallBusiness2	isecondisadvwomenownedsmallbusiness
167	vendor_vendorSiteDeta__ors_isJointVentureWomenOwnedSmallBusiness	isjointventurewomenownedsmallbusiness
168	vendor_vendorSiteDeta__s_isNativeHawaiianOwnedOrganizationOrFirm	isnativehawaiianownedorganizationorfirm
169	vendor_vendorSiteDeta__erviceRelatedDisabledVeteranOwnedBusiness	srdvobflag
170	vendor_vendorSiteDeta__cioEconomicIndicators_isTriballyOwnedFirm	istriballyownedfirm
171	vendor_vendorSiteDeta__dorSocioEconomicIndicators_isVeteranOwned	veteranownedflag
172	vendor_vendorSiteDeta__endorSocioEconomicIndicators_isWomenOwned	womenownedflag
173	vendor_vendorSiteDeta__nomicIndicators_isWomenOwnedSmallBusiness	iswomenownedsmallbusiness
174	vendor_vendorSiteDeta__Owned_isAsianPacificAmericanOwnedBusiness	apaobflag
175	vendor_vendorSiteDeta__inorityOwned_isBlackAmericanOwnedBusiness	baobflag
176	vendor_vendorSiteDeta__rityOwned_isHispanicAmericanOwnedBusiness	haobflag
177	vendor_vendorSiteDeta__cIndicators_minorityOwned_isMinorityOwned	minorityownedbusinessflag
178	vendor_vendorSiteDeta__norityOwned_isNativeAmericanOwnedBusiness	naobflag
179	vendor_vendorSiteDeta__cators_minorityOwned_isOtherMinorityOwned	isotherminorityowned
180	vendor_vendorSiteDeta___isSubContinentAsianAmericanOwnedBusiness	saaobflag

(a) Mapping of common attributes

	Unique Attributes	
	Unique Attributes in fpds	Unique Attributes in PrimeContracts
1	competition_idvTypeOfSetAside	congressionaldistrict
2	competition_numberOfOffersReceived	divisionnumberorofficecode
3	competition_numberOfOffersSource	emergingsmallbusinessflag
4	competition_typeOfSetAsideSource	fiscal_year
5	contractData_inherentlyGovernmentalFunction	hubzoneflag
6	contractData_listOfTreasuryAccounts_treasuryAccount_initiative	isarchitectureandengineering
7	contractData_listOfTr__yAccounts_treasuryAccount_obligatedAmount	isconstructionfirm
8	contractData_listOfTr__nt_treasuryAccountSymbol_agencyIdentifier	isotherbusinessororganization
9	contractData_listOfTr__unt_treasuryAccountSymbol_mainAccountCode	isserviceprovider
10	contractData_listOfTr__ount_treasuryAccountSymbol_subAccountCode	lastdatetoorder
11	contractData_undefinitizedAction	lettercontract
12	contractMarketingData_feePaidForUseOfService	locationcode
13	legislativeMandates_constructionWageRateRequirements	maj_agency_cat
14	legislativeMandates_laborStandards	maj_fund_agency_cat
15	legislativeMandates_l__lReportingValues_additionalReportingValue	mod_agency
16	legislativeMandates_materialsSuppliesArticlesEquipment	mod_parent
17	transactionInformation_closedBy	multipleorsingleawardidc
18	transactionInformation_closedDate	parentdunsnumber
19	transactionInformation_closedStatus	pop_cd
20	transactionInformation_createdBy	prime_awardee_executive1
21	transactionInformation_createdDate	prime_awardee_executive1_compensation
22	**transactionInformation**_lastModifiedBy	prime_awardee_executive2
23	vendor_vendorHeader_vendorAlternateName	prime_awardee_executive2_compensation
24	vendor_vendorSiteDeta__rtifications_isSBACertified8AJointVenture	prime_awardee_executive3
25	vendor_vendorSiteDeta__endorCertifications_isSBACertifiedHUBZone	prime_awardee_executive3_compensation
26	vendor_vendorSiteDeta__ations_isSelfCertifiedHUBZoneJointVenture	prime_awardee_executive4
27	vendor_vendorSiteDetails_vendorDUNSInformation_cageCode	prime_awardee_executive4_compensation
28	vendor_vendorSiteDeta__rganizationFactors_**countryOfIncorporation**	prime_awardee_executive5
29	vendor_vendorSiteDeta__rOrganizationFactors_stateOfIncorporation	prime_awardee_executive5_compensation
30	vendor_vendorSiteDeta__cioEconomicIndicators_isVerySmallBusiness	programacronym
31		progsourceaccount
32		progsourceagency
33		progsourcesubacct
34		psc_cat
35		rec_flag
36		statecode
37		streetaddress3
38		typeofidc
39		unique_transaction_id
40		vendorenabled
41		vendorlocationdisableflag

(b) Unique attributes of each table

Quality Assessment on Common Identity Attributes

Identity attributes play a critical role in identifying key entities of a contract. Quality issues on these attributes are not only usually hard to be resolved, and they often cause invalid data. This study mainly focuses on the quality assessment of key identity attributes on two dimensions: column completeness, and field length consistency, because the assessment of these dimensions doesn't require the knowledge of gold standard of data.

Completeness can be measured in different aspects including column completeness, schema completeness, and population completeness. Column completeness measures the degree to which there exist missing values in a column of a

table. Schema completeness measures the degree to which entities and attributes are missing from the schema. Population completeness measures the degree to which members of the population that should be present but not present. Since there is not enough information for assessing schema and population completeness, the study will focus only on column completeness, which is measured by the percentage of non-missing values in the column.

Field length consistency measures how consistent are lengths of attribute values. Most of identity attribute are supposed to have fixed-length values. For example, DUNS number, provided by Dun & Bradstreet (D&B), is a unique nine digit identification number for each physical location of a business. Thus a DUNs number of length other than nine-digit long is problematic. Although there are more than one identity attributes for each site, only assessment result on the key common attributes of both sites, including prime award ID, prime contractor DUNs, and contract agency ID, is reported here.

Table 3 shows that FPDS table has a higher column completeness measure than PrimeContracts table. Figures 2 to 4 show the field length distribution of PIID(prime award ID), prime contractor DUNS numbers, and contract agency ID respectively. Since PIID is a system wide identifier for each prime award, it is assumed to have a fixed length. So is contract agency ID. But there are some exceptions in both FPDS and primeContract tables. Similarly, DUNs number is a 9-digit value. Any DUNS numbers other than 9-digit are considered incorrect.

Table 3: Column completeness

Table Name	ColCnt	IncompleteCols	%CompleteCols
PrimeContracts	212	162	23.6%
FPDS	210	136	35.2%

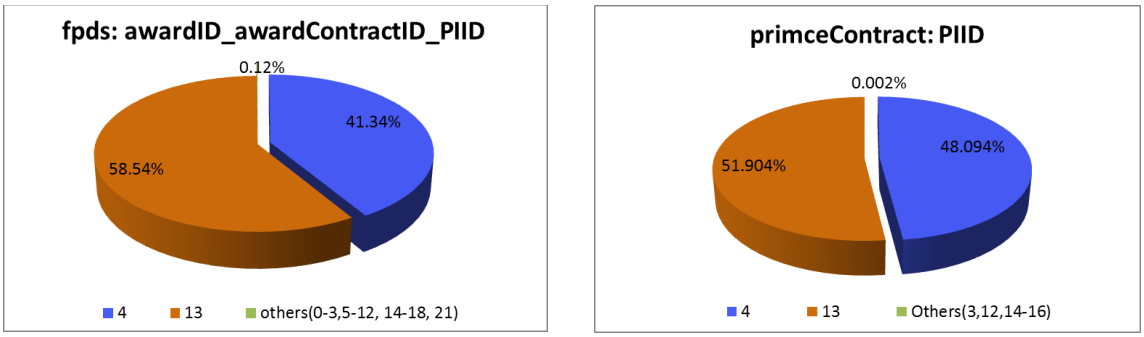


Figure 2: PIID Length Distribution

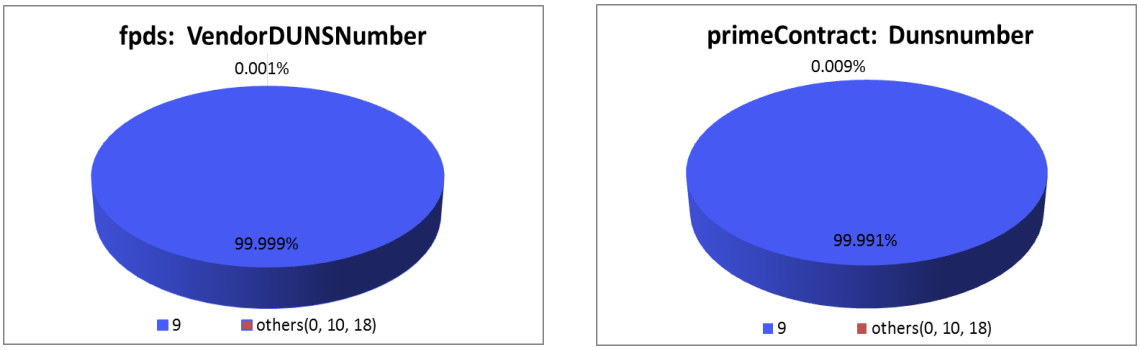


Figure 3: DUNs Number Length Distribution

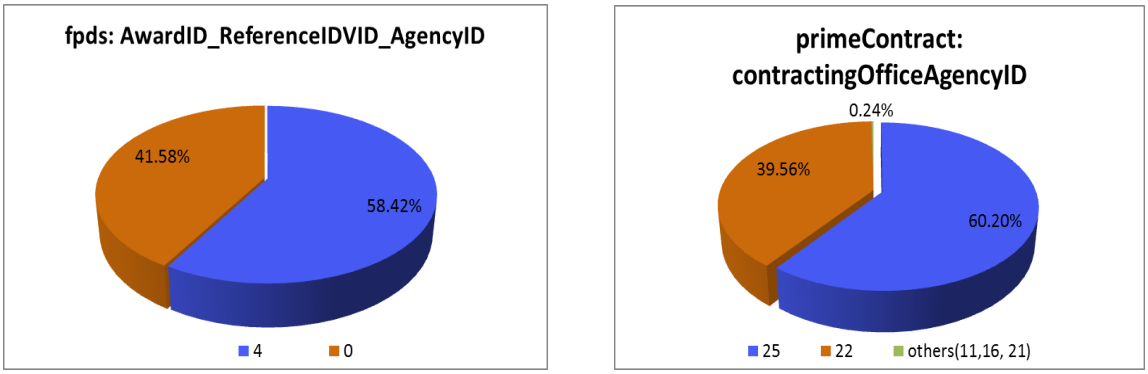


Figure 4: Contract Agency ID Length Distribution

Record mapping

Record mapping matches records of two tables if they represent the same entity. In FPDS and PrimeContracts, each contract is considered as an entity. Since both tables contain the contract information from the Department of Defense, record mapping provides a way for measuring the data consistency between the two. Record mapping is a typical entity resolution process. It requires comparing fields of records to determine whether they belong to the same entity or not. If records have common key identifier attributes, mapping them is rather straightforward; otherwise, the non-identifier

attributes have to be used to determine how similar the records are. Unfortunately, FPDS and PrimeContracts tables don't have a common record identifier, thus record mapping has to rely on the common attributes of two tables.

Considering the number of attributes and records in FPDS and PrimeContracts tables, record mapping is a very complicated and time-consuming process. Thus, the first phase of mapping is performed on sample data instead, and it only considers the following identity attributes when matching records: PIID, dunsnumber, vnedorlocationzipcode, vendorlocationstate, vendorlocationcity, vendor_countrycode, vendor_phoneno and vendorlocation_streetaddress. Here, PIID denotes the primary project ID that is unique to each project. Dunsnumer denotes the 9-digit DUNS number of the primary contractor of a project. vendorlocationzipcode, vendorlocationstate, vendorlocationcity, vendor_countrycode, vendor_phoneno and vendorlocation_streetaddress represent address and telephone information of a primary contractor. Two records are considered representing a same entity if their values on each of the above attributes match.

The following steps are performed to prepare the sample datasets.

- A random sample of 5000 common PIIDs that exist in both tables is drawn.
- The corresponding records of these PIIDs are retrieved from FPDS and primeContract tables respectively, and they are stored into separate datasets, denoted as datasets Df, and Du.
- As data quality issues will adversely affect the record matching result, data standardization and transformation are performed. Duplicate records are removed, and records with missing values are removed as well.
- The equijoin is applied on two datasets, and the resulting dataset is denoted as Djoin.

Figure 5 compares the number of distinct values of each identity attribute among three datasets Df, Du, and Djoin. It shows that Du consistently has more distinct values for each attribute than Df. The number of distinct values for each attribute in table Djoin indicates the number of common attribute values between Df and Du.

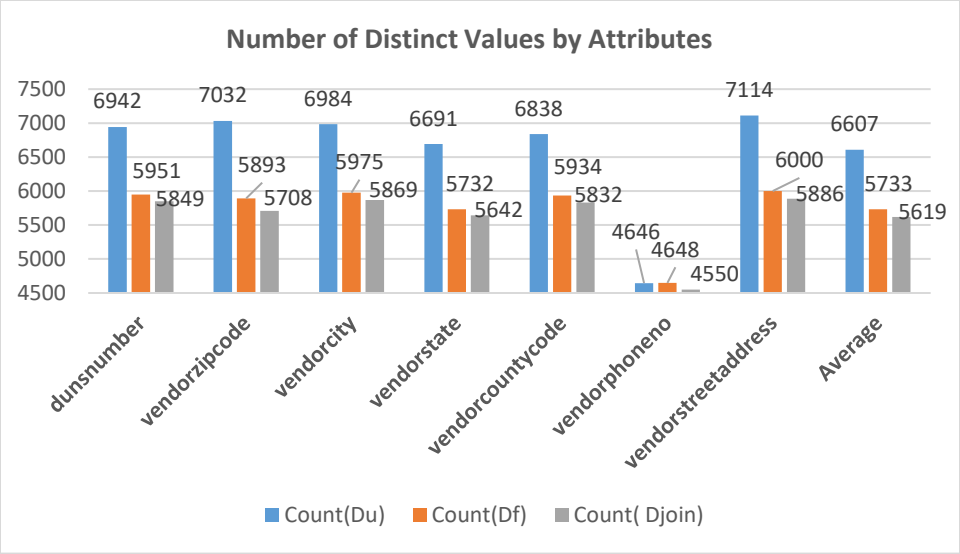


Figure 5: Number of distinct values by attributes

Figure 6 shows the relative consistency measure of each attribute of one table in terms of the other table. For example, 98.3% of dunsnumbers in Df are also found in Du, while only 84.3% of dunsnumber in Du are found in Df; 96.7% of vendorzipcodes in Df are also found in Du, but 81.2% of vendorzipcodes in Du are found in Df. The reason behind these discrepancies is that, given a prime award ID, there are more distinct records in Du than in Df. Possible root causes may include: FPDS.gov and usaspending.gov collected the data at different granularity levels, FPDS database may miss some records, or usaspending database may need to keep multiple records for the same prime award because these records are different.

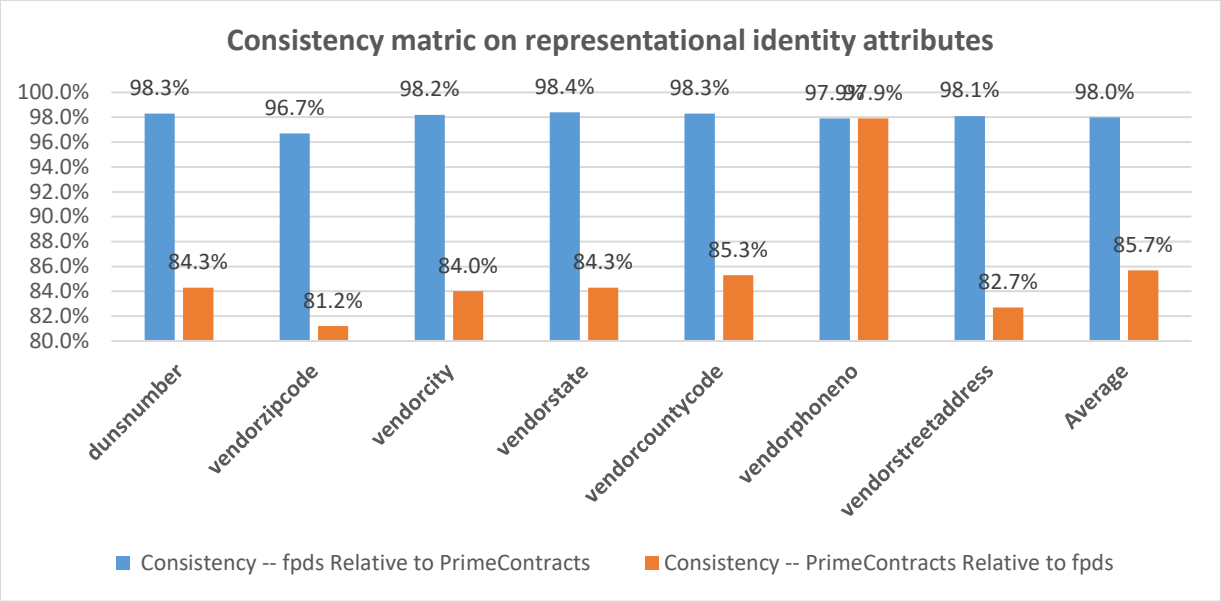


Figure 6: Relative consistency measure of each attribute

THIS PAGE LEFT INTENTIONALLY BLANK

Data Analytics

The goal of data analytics is to discover hidden and interesting patterns that can be potentially useful in planning future acquisition projects. The purpose of the project was to complement the expertise of domain experts on acquisition data and policies by pursuing a data science approach on multiple tracks.

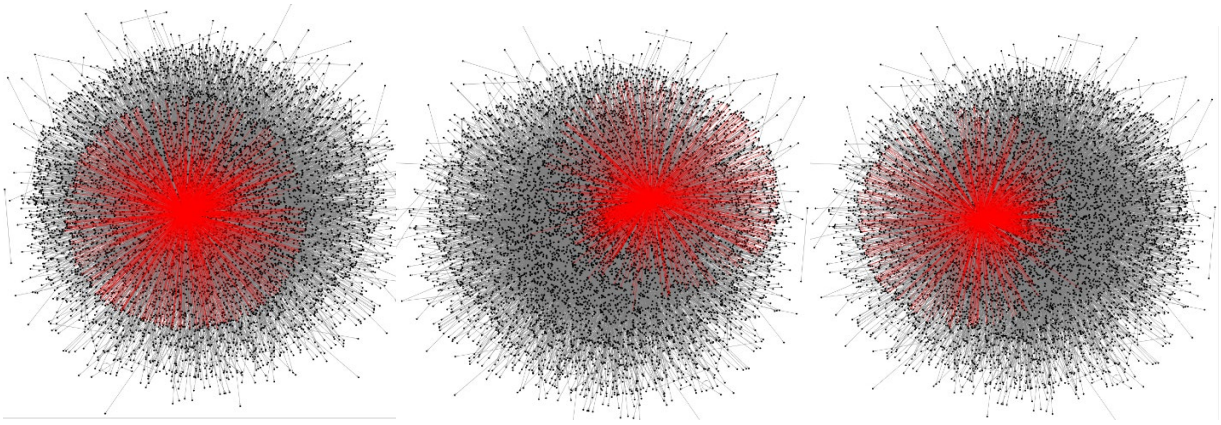
Data Mining Track

Data mining is the process of examining large data sets to uncover hidden but interesting patterns such as unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can shed significant insights to help add perspective to use the data and to lead to more effective decision makings. Some major data mining techniques include association discovery, classification, clustering, regression, sequence or path analysis, and structure and network analysis.

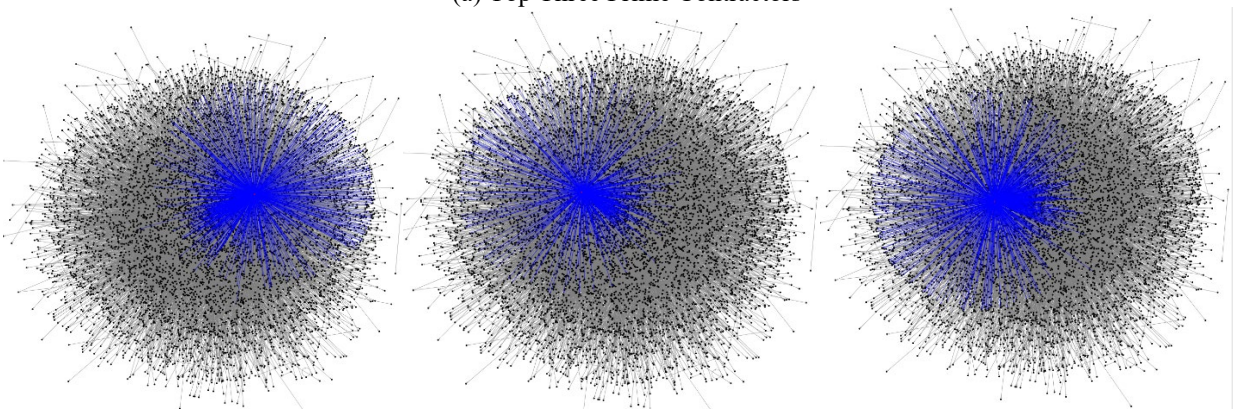
Association discovery aims to find frequent patterns that represent the inherent regularities in the datasets. Applications of association discovery include association, correlation, and causality analysis, basket data analysis, cross-marketing, etc. Classification, also called supervised learning, is the task of inferring a function from labeled training dataset. The function can then be used to classify new data instances. Decision tree, Bayesian networks, support vector machine, and neural networks are some of the commonly used models for classification. Clustering, also called non-supervised learning, group a collection of data objects into groups according a predefined distance function. Clustering can be employed as a stand-alone tool to get insights about data or as a preprocessing tool for other algorithms. Sequence analysis discovers patterns among sequences of ordered events or elements. Application of sequence patterns include customer shopping sequence, DNA sequences and gene structures, sequences of stock market changes, etc. Graph and network analysis aims to discover frequent subgraphs, trees, or substructures. It has been used for social networks analysis and web mining.

Cluster and Network Analysis

As the first phase of the research, network analysis is performed on prime contractors and their subcontractors of the sample database in a hope to discover the business networks among contractors. Figure 7 visualizes some findings from the network analysis. It shows the top three big contractors that have the largest number of subcontractors, and top three highly demanded subcontractors who are working for the largest number of different contractors.



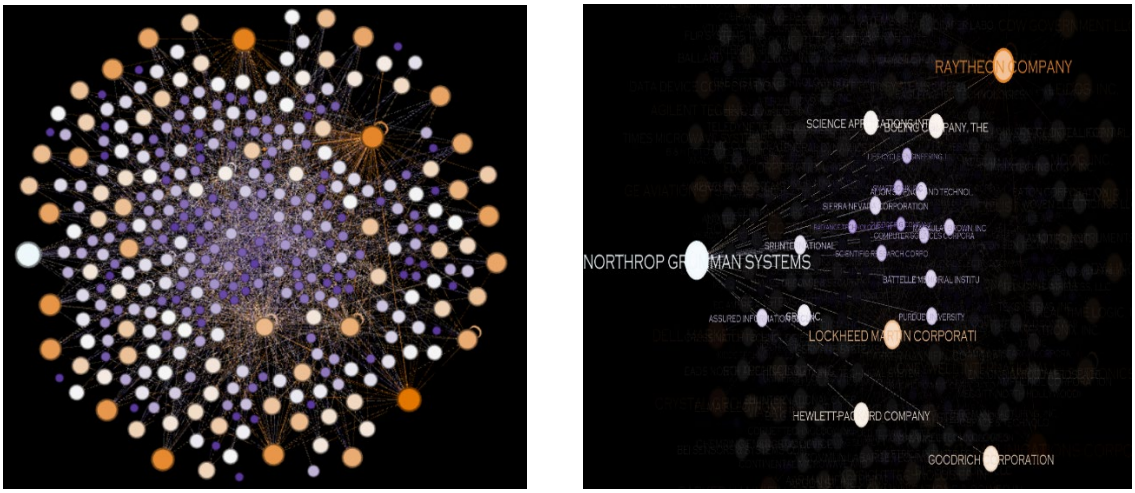
(a) Top Three Prime Contractors



(b) Top Three Subcontractors

Figure 7: Cluster Analysis of Contractors

Figure 8 shows the clustering results of only contractors that worked with at least 5 subcontractors. Figure 8(a) shows overall clustering result, where each dot represents a primary contractor. The dots in orange are “big” primary contractors with many subcontractors. The dots in purple are relatively “small” primary contractors. Figure 8(b) shows zoomed-in clusters for two big prime contractors.



(a) (b)

Figure 8: Clustering results for contractors involved in more than 5 projects

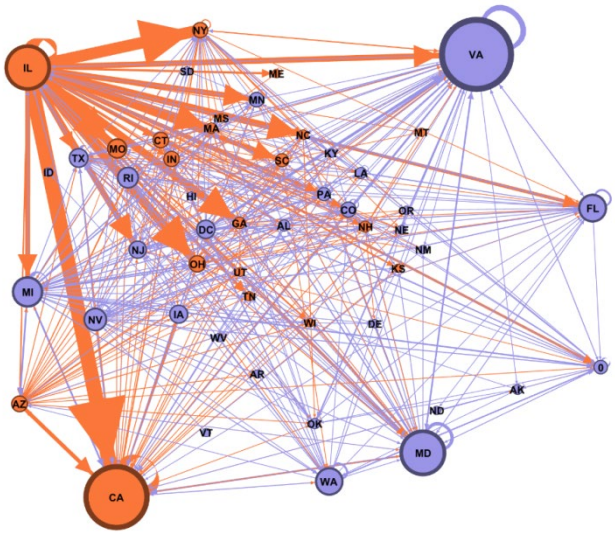


Figure 9: Clustering results by State

Figure 9 shows the analysis of the relationship between contractors and subcontractors by State. Each dot represents a state. The size of a dot is determined by the number of contracts awarded to a state. A directed edge represents the relationship between primary contractors and their subcontractors (pointed by an arrow). The thicker an edge is, the more contracts are between the primary contractors and their subcontractors. The figure shows some states, like California, get more contracts than others, and some states, like Illinois, tend to subcontract their projects to the other states.

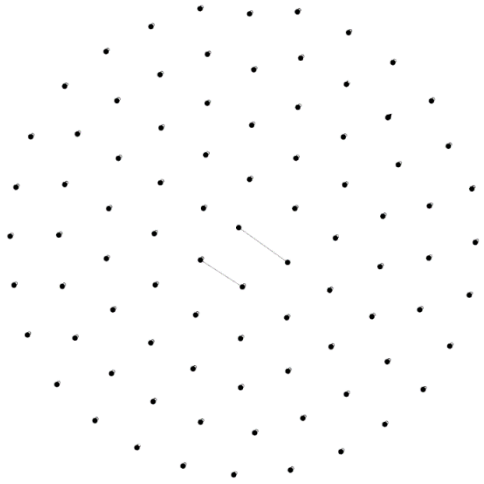


Figure 10: Relationship of companies of different business types

Figure 10 shows relationship of contractors of different business types. Each dot represents contractors of the same business type. A line between two dots indicates companies of two different business types are related by a contract. The figure reveals that companies tend to give subcontracts to the companies of the same business type. There are only two outliers that relate companies of different types.

Pattern Discovery

A preliminary data analysis was performed and aimed to discovery patterns that may shed insights into possible areas for improvement in acquisition projects. For instance, small contractors are usually less robust and easier to fail compared to large

contractors when facing natural or man-made disasters. Projects with subcontractors located in places that have a high risk of natural disasters such as earthquakes may have risks of a potential delay in delivery time. By taking into account of the risk factors in planning a project can help identify the room for improvement to ensure the successful and prompt delivery of the project.

The first round of analysis focused on finding the following patterns in the existing projects: 1) small-business subcontractors that are involved in different projects led by some key primary contractors; and 2) projects that have multiple subcontractors located in a place that has a high risk of natural disasters such as earthquakes, hurricanes, flooding, wild fires, etc. The following section discussed two examples of our findings.

PTB is a small, single-location company with less than 200 employees. It was involved in 6 different projects led by some key primary contractors including Boeing, Lockheed Martin, and L-3 Communications. The average award amount is about $5400. A close study on the company's website, shown as Figure 11, revealed it may provide some important services to its primary contractors. Since company websites and acquisition database are all publicly accessible, they might be used by enemies for inferring sensitive information on a project or planning attacks to make the project fail.

Figure 11: Snapshot of PTB Webpages

Critical Contractor Track

We define critical contractors as those that provide unique products and services. They could be the weakest link in a supply chain, because if they failed, it would be hard to find alternatives to fill their places.

NAICS (North American Industry Classification System) is the standard used by Federal statistical agencies in classifying business establishments for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy. NAICS code describes the business specialization of a company.

There are 379 distinct NAICS codes among all contractors in our databases. 78 NAICS codes have only one contractor associated with it. This means in the current pool of DoD contractors, these 78 contractors are critical contractors as no other DoD contractors are doing the same business as them. It is possible that there are companies that, outside the DoD contractor pool, have the same NAICS codes as those contractors. On average each of those critical contractors is involved in 37 different projects. The top ten critical contractors, which are involved in the most number of projects, are listed in Table 4.

Table 4: Top ten critical contractors with the most number of projects

Rank	No. of Distinct Projects
1	399
2	382
3	343
4	245
5	237
6	138
7	117
8	91
9	69
10	61

For those highly demanded contractors, most of them are big and well-established companies, but a couple of them are small companies that appear to provide very unique products and services. These companies could be a potential weak point in a project/supply chain and may critically affected the overall outcome of a project if they fail.

Exposure to Natural Disasters Track

A primary award usually has hundreds of contractors working on it. These contractors spread out in different geographical locations. Some may locate in an area with a high risk of natural disasters such as earth quake, flooding, hurricane, tornado, etc. Some natural disasters like tornado and earth quake are hard to predict. So it would be always beneficial to consider those risk factors when planning a project. Possible strategies include using contractors located in low risk areas, intentionally selecting contractors that are spread out in different geographical locations, or having backup plans in place to handle any emergencies.

As a proof of concept, the project initially focused narrowly on earthquakes only. Locations of 7.0-magnitude quake epicenters in US from U.S. Geological Survey website www.usgs.gov are retrieved. 118 subcontractors in SubContracts table are found located nearby an epicenter. Here are some findings:

- 984 awards have at least one subcontractor located in the high risk areas.
- 41 of them have at least two subcontractors located in the high risk areas.

The table below shows the top five contracts with the most number of subcontractors in high risk earth quake areas.

Table 5: Top 5 contracts with the most number of subcontractors in high risk earthquake areas

Project ID	#Subcontractors
1	15
2	15
3	11
4	10
5	8

In a more in-depth analysis, we have obtained the natural disaster data for each US county between the years 1950 and 2018 from the National Centers for Environmental Information (Formerly the National Climatic Data Center NCDC). The data cover all types of natural disasters, including flood, tornado, hurricane, blizzard, high wind, flash flood, hail, dust storm, etc.

This project focuses on disasters that could cause severe damages and significantly affect the normal life and business operations of local communities such as Tornado, Hurricane, Flood, and wildfire. Since the world weather has changed quite fast in recent decades, we decided to use the NCDC data of last twenty years to identify whether an area is prone to a natural disaster based on the following criteria. The high-risk flooding areas are identified as those that have at least 10 episodes of flood in the last twenty years; the high risk hurricane areas are those that have at least one hurricane in last 20 years; the high-risk wildfire areas are those that have at least one wildfire that lasted more than 1 day in last 20 years; and the high-risk tornado areas are those that have at least one category 3 or above tornado in the last 20 years. Table 6 shows the number of subcontractor zip codes belong to each disaster type:

Table 6: Number of subcontractor zips vulnerable to each disaster type

Disater type	Flood	Hurricane	Tornado	Wildfire
# zipcodes	5959	780	1182	1831

Our analysis found that there are 6786 natural-disaster-prone zip codes where the work of at least one subcontract was performed. Some of these zip codes are vulnerable to more than one disaster type. The natural-disaster-prone areas are further categorized into four classes based on the number of distinct disaster types that has been observed in that area during the last twenty years.

Table 7 shows the distribution of subcontract principal place zip codes by the number of disaster types along with the distribution of subcontractors located in those zip codes. The column %zip_population indicates the percentage of zip codes (of a category) with regard to the total number of subcontract zip codes, and %DUNS_population indicates the percentage of DUNs in each category of zips with regard to total number of subcontractor DUNS number. For instance, there are 2165 distinct zipcodes are vulnerable to one disaster type. These zipcodes account for 7.8% of all contractor zipcodes, and 42.3% of contractors located in these areas. There are 69 zipcodes that are vulnerable to all four disasters. These zipcodes account for 0.25% of all contractor zipcodes, and about 0.44% of contractors located in these areas.

Table 7: Distribution of subcontractor principal zip and DUNS

#DisasterTypes	#zipcodes	%zip_population	#duns	% DUNS_population
1	2165	7.8%	13373	42.3%
2	3548	12.9%	10965	34.6%
3	1004	3.6%	2733	8.6%
4	69	0.25%	141	0.44%
Total:	6786	23.7%	27072	86.0%

Subcontractors that are located in an area vulnerable to all four disaster types are considered to have a high risk. Table 8 shows top 10 projects with the most number of high risk contractors.

Table 8: Top 10 projects with the most number of high-risk contractors

Rank	No. of High-risk Contractors
1	59
2	49
3	43
4	37
5	36
6	31
7	27
8	24
9	23
10	19

It would be interesting to know the percentage of high risk contractors in past projects. There are total 588 projects have at least one high risk subcontractors. Figure 12 shows the distribution of projects by their percentage of contractors that are vulnerable to all four types of natural disasters. A close study reveals that the majority of 129 projects in the last bin with more than 90% of subcontractors in high risk areas have only one subcontractor. More than half of 588 projects have less than 10 percent of subcontractors in high risk areas. Ideally, a project should have as few as possible high-risk subcontractors.

We believe the information on high risk areas of natural disasters is beneficial as it helps project managers calculate the risk of a project and develop strategies to mitigate the risk to the minimum.

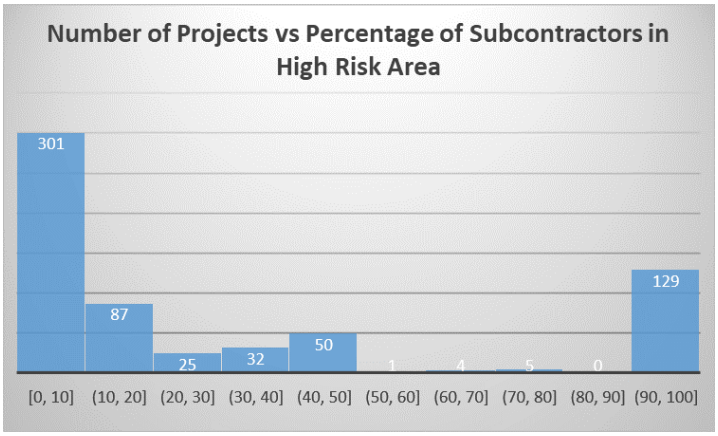


Figure 12: Distribution of projects by percentage of high-risk subcontractors

Natural Disaster Risk Map for U.S. Counties Track

National Centers for Environmental Information (NCEI) has been collecting natural disaster data since 1950 and has that information available for each U.S. county. The data covers a wide range of natural disasters, including flood, tornado, hurricane, blizzard, high wind, flash flood, hail, dust storm, etc. Our previous analysis on NCEI data revealed that it is challenging to categorize a disaster by its intensity and damage level because disasters lack such a categorization system. Even though some disaster types, such as tornadoes and hurricanes, do have a categorization system, it is often difficult to assess an incident's impact to the local communities without other supporting information. Furthermore, a comparison of impact of disasters of different types is not easy to perform.

One objective of this research was to identify U.S. areas with a high risk of natural disasters. To assess an area's risk level, we considered both the number of disasters and the impact of those disasters. For example, an area might have experienced several minor natural disasters during the period we studied, but none of them was serious, while another area has fewer incidents, but some of them were serious and had a serious effect on local communities and economy. For the purpose of acquisition risk management, the second area in our example should be considered as a high risk zone.

To gain a better understanding of a disaster in terms of its intensity and impact, NCEI data is enhanced with FEMA disaster mitigation and recovery data. Table 9 lists the FEMA data fields:

Table 9: FEMA data fields

1	femaDeclarationString	13	incidentBeginDate
2	disasterNumber	14	incidentEndDate
3	state	15	disasterCloseoutDate
4	declarationType	16	fipsStateCode
5	declarationDate	17	fipsCountyCode
6	fyDeclared	18	placeCode
7	incidentType	19	designatedArea
8	declarationTitle	20	declarationRequestNumber
9	ihProgramDeclared	21	hash
10	iaProgramDeclared	22	lastRefresh
11	paProgramDeclared	23	id
12	hmProgramDeclared		

FEMA data shows the beginning and ending dates of an incident, its location information, and the FEMA assistance program declared. The types of assistance programs actually provide a good indicator on the damage level and scope of an incident. Below are their short descriptions:

- ihProgramDeclared: denotes whether the Individuals and Households program was declared for this disaster.
- iaPorgramDeclared: denotes whether the Individual Assistance program was declared for this disaster
- paProgramDeclared: denotes whether the Public Assistance program was declared for this disaster
- hmProgramDeclared: denotes whether the Hazard Mitigation program was declared for this disaster.

Among the four assistance programs, ihProgram is the highest level of assistance and aims to help communities that are significantly affected by a major disaster, and hmProgram is the lowest. Our proposed approach is to use the declared assistance programs to assess a natural disaster's intensity and damage level. More specifically, given the number of disasters in an area during a period and the corresponding number of different assistance programs declared, a **weighted sum of disaster** number, hence termed weighted disaster score (WDS), is calculated as the follows:

$$s = \sum_{i=1}^{4} w_i \times n_i$$

Where n_i is the number of a specific type of the assistance programs, and w_i is the corresponding weight for the type. The weight for each assistance program is defined as follows:

- Disaster mitigation: 0.25
- Public assistance : 0.50
- Housing assistance: 0.75
- Individual assistance: 1.0

Table 10 shows the five-number summary (i.e., minimum, first quartile (Q1), medium, third quartile (Q3), and maximum values) for the WDS scores and the number of disasters of all U.S. counties between 1953 and 2020, respectively.

Table 10: Five-number summary of WDS and number of disasters of FEMA data

Statistic	WDS	# of Disasters
Min	0.25	1
Q1	6.25	10
Medium	10	15
Q3	13	19
Max	56.25	105

To facilitate comprehension for a wide range of domain experts and program officers, we define three risk levels, namely low, medium, and high. The first quartile and the third quartile of WDS are used as the cutoff points for the risk classes as shown in Table 11. A natural disaster risk class is assigned to each county based on its WDS value. Counties with a WDS value less than 6.25 are considered to have low risk of natural disasters; counties with a WDS value in between 6.25 and 13 are considered to have medium risk; and counties with a WDS value greater than 13 are considered to have high risk.

Table 11: Risk level Categorization

level	WDN values
Low	wdn<1st quartile
Medium	1st quartile <wdn<3rd quartile
High	wdn> 3rd quartile

Figure 13 shows the distribution of U.S. counties by risk levels. It shows that about 28% of counties are located in a high-risk area, 52% in medium risk, and 20% in low risk. Figure 14 shows the disaster risk class on a geographical map of the U.S.

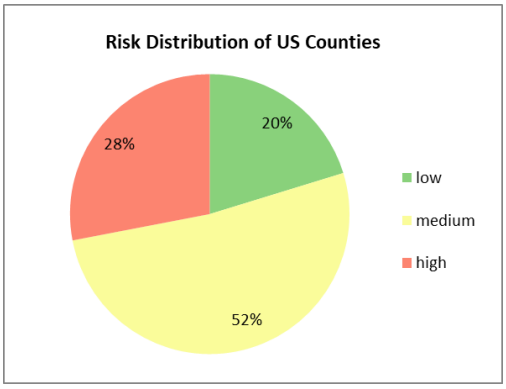


Figure 13: County Distribution by Risk Levels

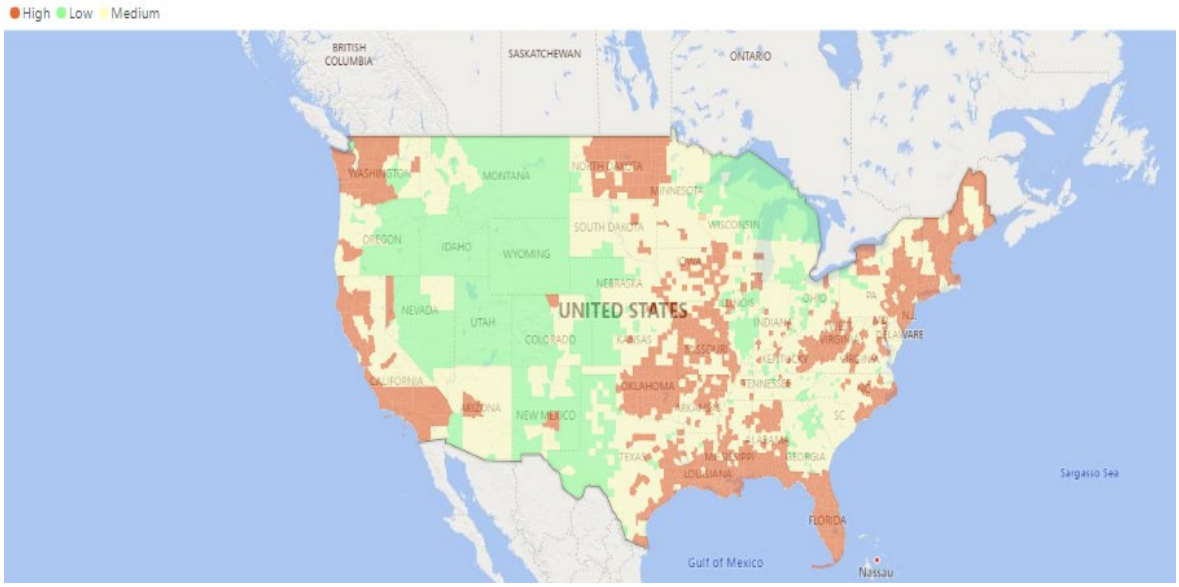


Figure 14: Natural disaster risk class (low, medium, high) displayed for each county in the United States. The index takes into count both the number of occurrences of disasters and their magnitude. Red encodes high risk, yellow medium, and green low.

Distribution of Federal Contractors and Business in Different Risk Areas

Not all counties in the U.S. have federal contractors or business being conducted for a federal project, and the focus of this subsection is to investigate how various typed of Navy industries relate to disaster areas. To this end, we consider information from FPDS.gov and usaspend.gov and join it with the data from NCEI and FEMA. Figure 15

shows the distribution of federal contractors divided by class of natural disaster risk. It shows as high as 41% of contractors are located in high-risk areas. There are about 17% of contractors are not located in U.S. so their natural disaster risk levels are not assessed by this research. Figure 16 provides somewhat of a reverse analysis, where each county in the U.S. is mapped based on the risk type and number of awards received from the Navy. A map such as this can be made interactive and computed in real-time to show the distribution of any acquisition project and help a domain expert plan or run the project.

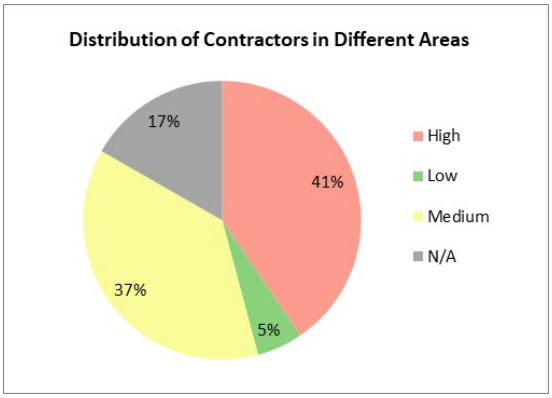


Figure 15: Contractors Distribution by Risk Levels. Not all business takes place in the in the U.S., and thus the risk level is not known for 17%.

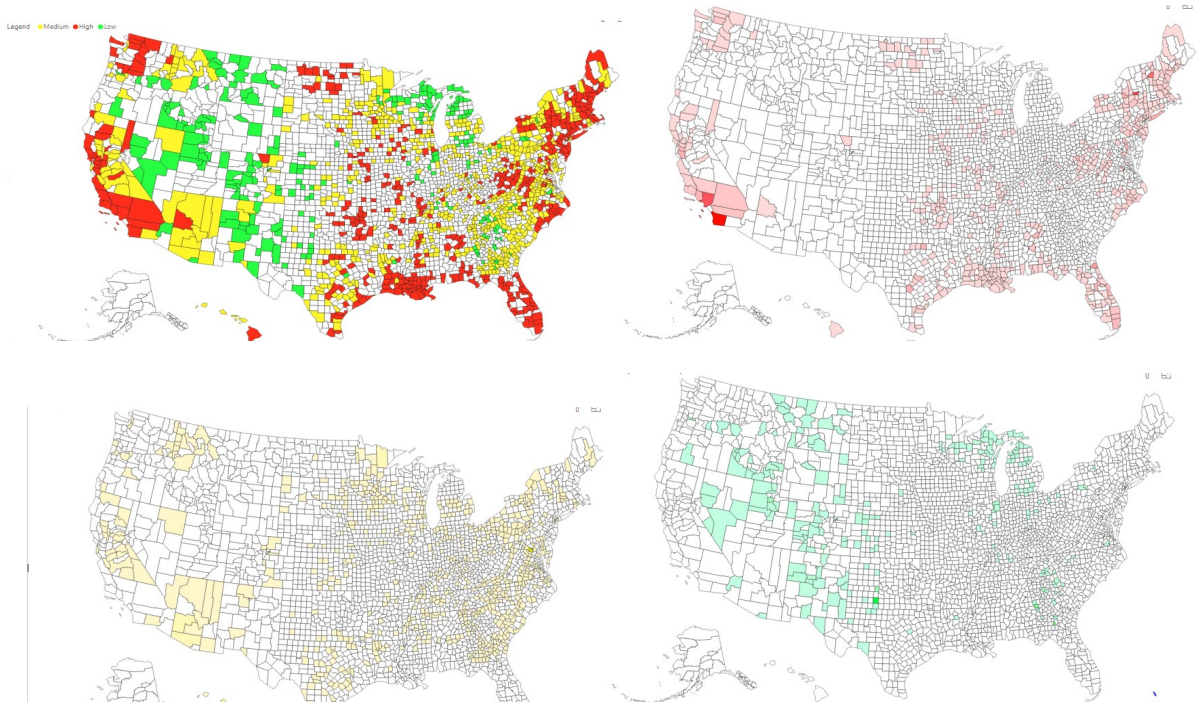


Figure 16: Place of performance of Navy awards correlated to natural disaster risk. Red is high risk, yellow is medium, and green is low Top-left: overall view of counties with at least one award. Top-right shows only counties with high risk using color intensity to encode the number of awards in that county. Bottom-left depicts counties with medium risk level, using color intensity for number of awards. Similarly, bottom-right shows low risk counties and number of awards.

North American Industry Classification System (NAICS, 2017) is the standard used by Federal statistical agencies in classifying business establishments for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy. NAICS code describes the business specialization of a company. There are 379 distinct NAICS codes among all contractors in our database copy of the federal purchasing information, and 355 of them have contractors located in natural disaster risk areas.

Generally, a specific type of industry would be robust and not susceptible to natural disasters if most of the companies providing services and products in that NAICS code are located in low risk areas. Figure 17 shows the clusters of NAICS codes based on the percentage of companies located in high-risk areas. Fifty two of the 355 NAICS codes have all companies doing business with the Navy in high-risk areas. A closer look at these 52 NAICS codes reveals that the majority of them have only one

contractor. Thus, companies of these NAICS code can increase the system risk of any acquisition project in which they participate.

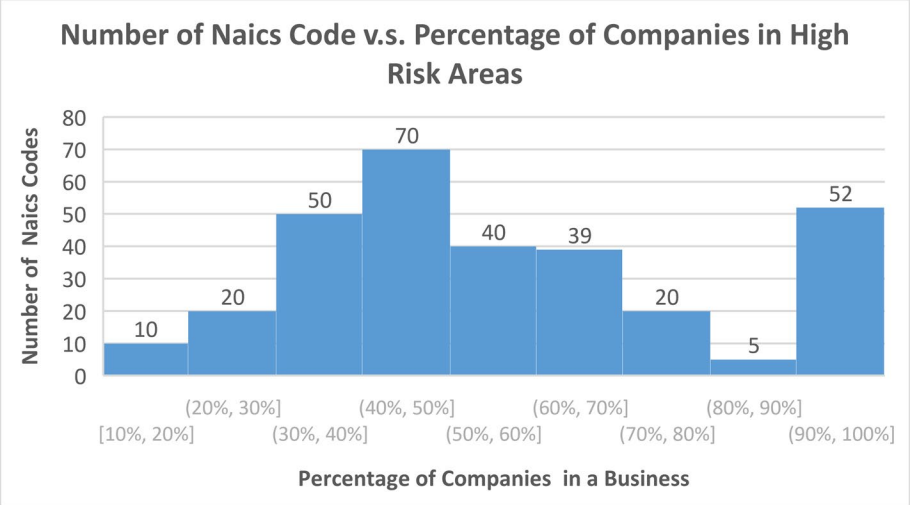


Figure 17: Clustering NAICS codes by percentage of high-risk companies

1.1.1 Distribution of High Risk Contractors by Project

This section analyzes the distribution of high-risk contractors, as defined by their NACS and location, in past federal projects. Considering USASpending data, there are about 13435 distinct projects with subcontractor information. The percentage of high-risk contractors for each project is calculated, and it reveals that more than half of the projects have high-risk contractors. A closer look at these projects shows that 90% are single contractor projects, and over 97% have no more three total contractors. The maximum number of contractors in these projects is 35. Figure 18 shows the number of projects in each percentage bin. It shows that except for the projects in the last bin the majority of the remaining projects have between 30% to 50% of high risk contractors.

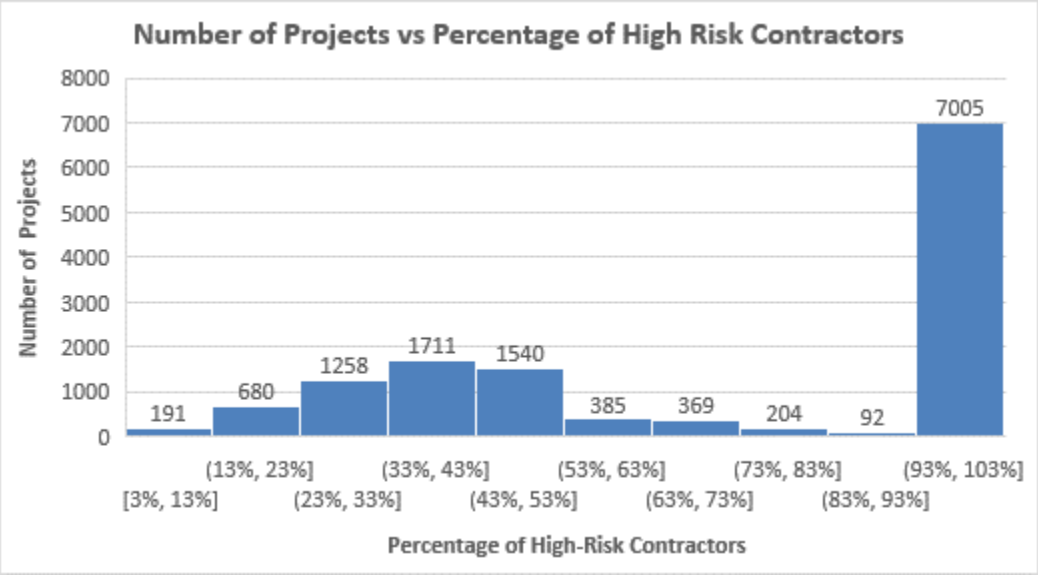


Figure 18: Clustering of projects by percentage of high-risk contractors

Because the number of contractors in a project varies, we analyze the distribution of number of contractors and present the results in Figure 19, which shows the scatter plot of the total number of contractors in a project and the percentage of the high-risk ones. It shows that the majority of projects have fewer than 250 contractors. There are several projects with more than 1000 contractors. Figures 20(a) and 8(b) partition the projects into two groups, the one with more than 100 contractors and the one with fewer than 100 contractors, and it shows the percentage of high-risk contractors for the projects in each group.

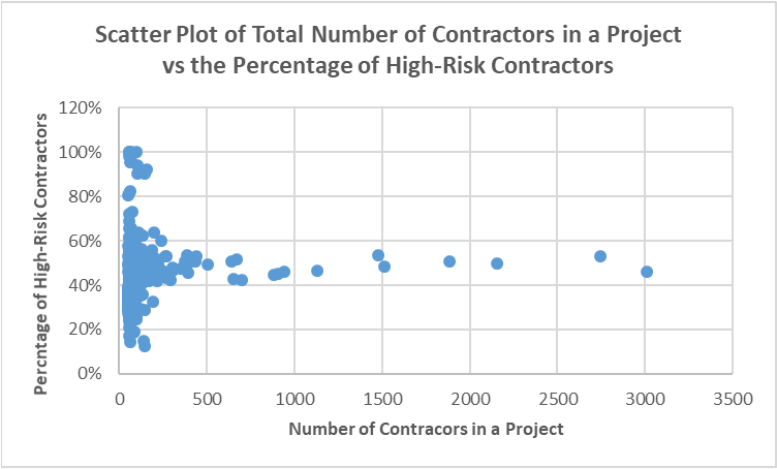
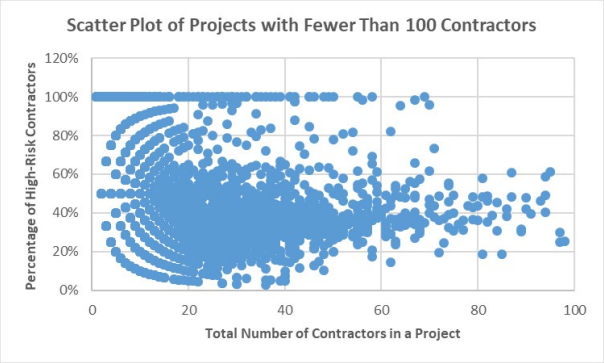
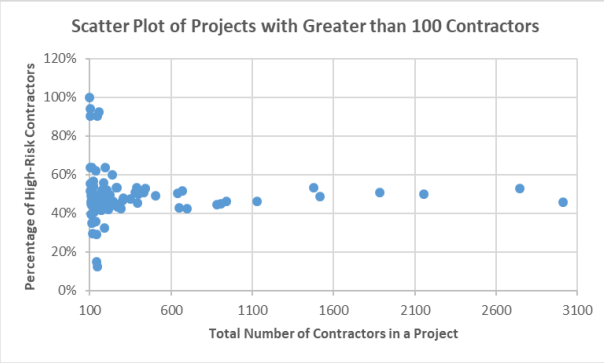


Figure 19: Percentage of high-risk contractors vs total number of contractors



(a)Fewer than 100 contractors per project



(b) Greater than 100 contractors per project

Figure 20: Percentage of high-risk contractors vs total number of contractors

Related Work

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 & Baney, 2012). To this end, research has looked at automatic means of dealing with the heterogeneous acquisition data sources from text processing (Zhao et al., 2015), systems engineering (Cilli et al., 2015), and business (Gaither, 2014) perspectives. Our research is different both in content and in the approach. In content in that we rely on big data to identify hidden risk factors, and in the approach in that our expertise in information visualization, data quality, data governance and policy (Chief Data Officers), and in data science provides a value-based perspective.

Tudoreanu et al investigated employment data in an attempt to correlate changes in employment with negative modifications to contracts (Tudoreanu et a. 2018). Such correlations can be explored to infer hidden and undisclosed contractors. Hidden contractors may pose the risk of becoming a weak, stress point of a project and would affect the overall outcome of the project.

Wu et al proposed a framework based on data science approach that aims to utilize the online information to assess and improve acquisition database quality as well as to find the hidden patterns to further acquisition research (Wu et al 2018). The main component of the framework is a web-search and text mining module, whose main function is to search the internet and identify the most credible and accurate information online.

Apte et al. made use of Big Data analytic techniques to explore and analyze large dataset that are used to capture information about DOD services acquisitions (Apet et al, 2015). The paper described how big data analytics could potentially be used in acquisition research. As the proof of concept, the paper tested the application of Big Data Analytic techniques by applying them to a dataset of CPARS (Contractor Performance Assessment Report System) ratings of 715 acquired services. It also created predictive models to explore the causes of failed services contracts. Since the dataset used in the research was rather small and far from the scope of big data, the

techniques explored by the paper mainly focus on traditional data mining techniques without taking into account of big data properties.

Black et al. studied the quality of narratives in Contract Performance Assessment Reporting System(CPARS) and their value to the acquisition process(Black et al, 2014). The research used statistical analysis to examine 715 Army service contractor performance reports in CPARS in order to understand three major questions: (1) To what degree are government contracting professionals submitting to CPARS contractor performance narratives in accordance with the guidelines provided in the CPARS user's manual? (2) What is the added value of the contractor performance narratives beyond the value of the objective scores for performance? (3) What is the statistical relationship between the sentiment contained in the narratives and the objective scores for contractor evaluations?

Conclusion and Future Work

This research presented a data science approach to compare and analyze publicly accessible acquisition databases. The research explored the usage of online information to enhance the internal data in order to discover hidden patterns in the data. The research has collected natural disaster information from the National Centers for Environmental Information. The information is helpful in identifying high risk locations and contractors located in those locations. This study considered only four disaster types. We plan to include more disaster types in our future study. As some disasters are correlated, such as hurricane and flooding, it would be interesting to identify the disaster types that are most damaging and disruptive to local business and categorize the disaster types accordingly based on their disruptive levels. Then a natural disaster risk model can be developed to assign a weighted risk factor for each zipcode based on its vulnerability to a particular disaster type.

Besides, our future work will cover the following two directions. First, explore more data analytics techniques to discover patterns that are potentially useful to acquisition research community. Second, research effective text mining techniques for assessing web data quality and retrieving credible information from online sources.

THIS PAGE LEFT INTENTIONALLY BLANK

References

Apte, U., Rendon, R., & Dixon, M. (2016). “Big Data Analysis of Contractor Performance Information for Service Acqusition in DoD: A Proof of Concept”. Proceedings of the Thirteen Annual Acquisition Research Symposium.

Augustine, Norman R. (1997) Augustine's laws. AIAA.

Black, S., Henley, J., & Clute, M. (2014). Determining the value of Contractor

Performance Assessment Reporting System (CPARS) narratives for the acquisition process (NPS-CM-14-022). Monterey, CA: Naval Postgraduate School.

Brown, Bradford. (2010) “Introduction to Defense Acquisitions Mgmt. 10th Ed., Defense Acquisition University.” www.dau.mil/publications/publicationsDocs/Intro%20to%20Def%20Acq%20Mgmt%2010%20ed.pdfCheskin, S. (1999). “Ecommerce trust: Building trust in digital environments,” Archetype/Sapient 1999.

Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002

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, 2015, pp. 584–603. doi:10.1002/sys.21329.

Corritore, C.L., Kracher,B., & Wiedenbeck,S. (2003) “On-line trust: Concepts, evolving themes, a model,” *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 737–758, 2003.

DAU. “DAU Center For Defense Acquisition Research Agenda 2016–2017.” http://dau.dodlive.mil/files/2016/01/ARJ-76_ONLINE-FULL.pdf.

Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., & Treinen, M. (2001) “What makes web sites credible?: A report on a large quantitative study,” in *CHI ’01: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 61–68, New York, NY, USA: ACM Press, 2001.

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

Gallup et al. (2015) “Lexical Link Analysis (LLA) Application: Improving Web Service to Defense Acquisition Visibility Environment.” Distributed Information Systems Experimentation, May.

Hagan, G. “Glossary: Defense Acquisition Acronyms and Terms.” Fort Belvoir, VA, Dept. of Defense, Defense Systems Management College, Acquisition Policy Dept., 1998.

Jennifer Golbeck (2008), "Trust on the World Wide Web: A Survey", Foundations and Trends® in Web Science: Vol. 1: No. 2, pp 131-197.

Krzysko, Mark. (2012) "The Need for Acquisition Visibility." Journal of Software Technology Feb.: 4-9.

Krzysko, M. (2016) “Acquisition Decision Making through Information and Data Management” www.digitalgovernment.com/media/Downloads/asset_upload_file917_5737.pdf.

McKernan, M., Moore, N.Y., Connor, K., Chenoweth, M.E., Drezner, J. A., Dryden, J., Grammich, C.A., Mele, J.D., Nelson, W., Orrie, R., Shontz, D., & Szafran, A. (2016). “Issues with Access to Acquisition and Information in the Department of Defense”. Technical report, Rand Corporation.

Metzger, M.J. & Flanagin, A.J. (2013). “Credibility and Trust of Information in Online Environments: The Use of Cognitive Heuristics”. Journal of Pragmatics 59 (2013), pp210-220.

Miller, A. & Ray, J. (2015). “Moving from Standard Practices to Best Practices in Defense Acquisition.” Defense ARJ, vol. 22, no. 1, 1 Jan. 2015, pp. 64–83.

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

Pennock, Michael J. (2008) “Defense Acquisition: A Tragedy of the Commons”. ProQuest,.

Tudoreanu, M.E., Franklin, K., Wu, N., & Wang, R. “Searching Hidden Links: Inferring Undisclosed Subcontractors from Public Contract Records and Employment Data,” *Proceedings of the Fifteenth Annual Acquisition Research Symposium*, May, 2018.

Undersecretary of Defense. (Nov 2007) “DoDI 5000.01 – Operation of the Defense Acquisition System”,

Undersecretary of Defense(2015) “DoDI 5000.02 – Operation of the Defense Acquisition System”,.

Wu, N., Tudoreanu, M.E., & Wang, R. (2018) “Leveraging Public Data for Quality Improvement and Pattern Discovery of Federal Acquisition Data*”, Proceedings of the Fifteenth Annual Acquisition Research Symposium*, May, 2018.

Acquisition Research Program
Graduate School of Defense Management
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net