

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO



CÓMO DETECTAR CORRUPCIÓN, COLUSIÓN Y FRAUDE  
EN LOS CONTRATOS QUE OTORGA EL BANCO MUNDIAL

T E S I S

QUE PARA OBTENER EL TÍTULO DE

MAESTRO EN CIENCIA DE DATOS

PRESENTA

CARLOS EDUARDO PETRICIOLI ARAIZA

ASESOR: DR. ADOLFO JAVIER DE UNÁNUE TISCAREÑO

México, D.F.

2016

*(This page is intentionally left blank)*

“Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada ‘CÓMO DETECTAR CORRUPCIÓN, COLUSIÓN Y FRAUDE EN LOS CONTRATOS QUE OTORGA EL BANCO MUNDIAL’ otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Baillères Jr., autorización para que fijen la obra en cualquier medio, incluido el electrónico y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación”.

CARLOS EDUARDO PETRICIOLI ARAIZA

---

FECHA

---

FIRMA

*Aquí van los agradecimientos*

*Aquí va la dedicatoria*

# Contents

---

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	The World Bank . . . . .	5
<b>2</b>	<b>The World Bank's Procurements</b>	<b>9</b>
2.1	The World Bank Group's Rules and Guidelines . . . . .	9
2.1.1	Procurements General Considerations . . . . .	10
2.1.2	Procurement Plan . . . . .	11
2.2	Corruption, Collusion and Fraud . . . . .	13
2.2.1	Examples: What is Fraud and Corruption? . . . . .	14
2.3	World Bank Listing of Ineligible Contractors . . . . .	16
2.4	Investigations . . . . .	19
2.4.1	Report Suspected Fraud or Corruption . . . . .	20
2.5	Searching for <i>red-flags</i> . . . . .	22
<b>3</b>	<b>The World Bank's Data</b>	<b>26</b>
3.1	Private data . . . . .	26
3.1.1	Investigations . . . . .	26
3.2	Public data . . . . .	29
3.2.1	World Development Indicators . . . . .	31
3.2.2	Major & Historic Awards . . . . .	39
3.3	Entity names disambiguation . . . . .	45

3.3.1	Entity names disambiguation parameters & results . . . . .	47
3.4	Co-Award Network and feature creation . . . . .	50
<b>4</b>	<b>Detecting corruption, collusion &amp; fraud: Data product</b>	<b>53</b>
4.1	Data pipeline . . . . .	53
4.2	Red-flags from data . . . . .	54
4.3	Model . . . . .	55
4.4	Web visualization application: dashboard . . . . .	55
4.4.1	Dashboard outline . . . . .	55
4.4.2	Interactive map . . . . .	55
4.4.3	Companies & projects network . . . . .	56
4.4.4	Contract specific risk map . . . . .	56
<b>A</b>	<b>Sample Procurement Plan</b>	<b>57</b>
<b>B</b>	<b>Software</b>	<b>61</b>
B.1	Amazon Web Service . . . . .	62
<b>C</b>	<b>World Bank countries</b>	<b>64</b>
C.1	World Bank’s World Development Indicators . . . . .	67
<b>D</b>	<b>Code</b>	<b>76</b>
D.1	Entity Names disambiguation code . . . . .	78
	<b>Bibliography</b>	<b>92</b>
	Books . . . . .	92
	Articles . . . . .	92
	Other references . . . . .	93

# List of Figures

---

2.1	Integrity Complaint Form . . . . .	21
2.2	What does corruption look like? . . . . .	24
3.1	WDI: % of Bribes to tax officials per Country/Region . . . . .	35
3.2	WDI: % of Firms competing against informal firms per country/Region	36
3.3	WDI: % of Payments to public officials per Country/Region . . . . .	37
3.4	WDI: Time to Enforce Contracts per Country/Region . . . . .	38
3.5	Number of Historic & Major awards per Region . . . . .	41
3.6	Total awarded amount of Historic & Major contracts per Region . . .	41
3.7	Top six countries: number of awarded contracts . . . . .	43
3.8	Top six countries: total awarded amount (\$USD) . . . . .	44
3.9	Co-award Network example . . . . .	51
4.1	Data pipeline . . . . .	53
4.2	Interactive map . . . . .	55
4.3	Risk map (Sample) . . . . .	56
C.1	WDI: Transparency Accountability Corruption Rating, Country/Region	67
C.2	WDI: Property Rights Rule Governance Rating per Country/Region .	68
C.3	WDI: Legal Rights Index per Country/Region . . . . .	69
C.4	WDI: % of Firms that do not report all sales per Country/Region . .	70



C.5	WDI: Business disclosure index per Country/Region . . . . .	71
C.6	WDI: Gross Domestic Product per Capita per Country/Region . . . .	72
C.7	WDI: Gini index of inequality per Country/Region . . . . .	73
C.8	WDI: % of Primary School graduation per Country/Region . . . . .	74
C.9	WDI: % of Unemployment per Country/Region . . . . .	75

# List of Tables

---

2.1	Debarred Firms and Individuals (sample) . . . . .	18
3.1	Major and Historic Awards Dictionary . . . . .	40
3.2	Entity names disambiguation example . . . . .	48
3.2	Entity names disambiguation example . . . . .	49
3.2	Entity names disambiguation example . . . . .	50
3.3	Co-award network features . . . . .	51
3.3	Co-award network features . . . . .	52
A.1	Prior Review Threshold . . . . .	58
A.2	Summary of the procurements packages . . . . .	59
A.3	Selection Method . . . . .	59
A.4	Consultancy Assignments with Selection Methods and Time Schedule	60

# CÓMO IDENTIFICAR CORRUPCIÓN, COLUSIÓN Y FRAUDE EN LOS CONTRATOS QUE OTORGA EL BANCO MUNDIAL<sup>1</sup>

CARLOS PETRICIOLI<sup>2</sup>

## Resumen

RESUMEN EN ESPAÑOL

**Keywords:** Aprendizaje de máquina, Corrupción, Fraude, El Banco Mundial.

---

<sup>1</sup> Última actualización: April 21, 2016.

<sup>2</sup> Agradezco especialmente a Rayid Ghani (Ph.D. Carnegie Mellon University) y a Eric Rozier (Ph.D. University of Illinois Urbana-Champaign) por su asesoría en la elaboración de este trabajo. También agradezco la colaboración y los comentarios de Jeff Alstott (Ph.D. Cambridge), Dylan Fitzpatrick (Ph.D. Carnegie Mellon), Misha Teplitskiy (Ph.D. University of Chicago) y de Elizabeth Wiramidjaja (Banco Mundial).

carpetri@gmail.com

detecting-corruption.carlospetricioli.com

# DETECTING CORRUPTION COLLUSION AND FRAUD<sup>3</sup>

## CARLOS PETRICIOLI<sup>4</sup>

### Abstract

The World Bank Group lends billions of dollars each year to fund development projects in its efforts to reduce global poverty. This project helps investigators at the Bank search for patterns of collusion, corruption, and fraud in its contracts data by using models of contract-specific risk. An automated approach like this can highly increase the World Bank's ability to detect these offenses by efficiently targeting their future investigations.

The actual problem is that contractors providing goods and services on World Bank's projects are typically hired through a competitive bidding process, but occasionally, prospective contractors influence the competitive system by colluding with other contractors, bribing government officials, or otherwise manipulating the bidding process. These offenses have far-reaching effects on the price and quality of contract delivery. This project develops contract-level risk models that use historical data on major contracts awarded from the past 20 years and internal investigations data, covering companies and projects investigated in the past.

To approach the problem, an interactive dashboard was developed for investigators to track any company's activity across countries, sectors, and time. By using this tool, investigators can track contract awards companies have received, including under different names, view a risk score for each World Bank contract, as calculated by our contract risk model and visualize the immediate neighborhood of the company in its co-award network.

In conclusion, current data is sufficient to forecast risk and allows investigators to be proactive in determining which companies, projects and contracts to examine.

**Keywords:** Machine Learning, Corruption, Fraud, The World Bank.

---

<sup>3</sup> Last Update: April 21, 2016.

<sup>4</sup> I especially thank Rayid Ghani (Ph.D. Carnegie Mellon University) and Eric Rozier (Ph.D. University of Illinois Urbana-Champaign) for their mentorship throughout the elaboration of this work. I also appreciate the collaboration and comments of Jeff Alstott (Ph.D. Cambridge), Dylan Fitzpatrick (Ph.D. Carnegie Mellon), Misha Teplitskiy (Ph.D. University of Chicago) and Elizabeth Wiramidjaja (World Bank).

carpetri@gmail.com

detecting-corruption.carlospetricioli.com

## Web page

---

[detecting-corruption.carlospetricioli.com](http://detecting-corruption.carlospetricioli.com)

---

## Chapter 1.

# Introduction

---

The World Bank Group lends billions of dollars each year to fund development projects in its efforts to reduce global poverty. This project helps investigators at the Bank search for patterns of collusion, corruption, and fraud in its contracts data, using models of contract-specific risk. By developing an automated approach to detect these offenses this project can help the World Bank efficiently target future investigations.

Several things are required for a project of this type to achieve success. The following list outlines the requirements for a successful project, some requirements easier to satisfy so it is ordered from easiest to most difficult. These alone do not guarantee success, but success is nearly impossible without them [web page *What Makes a Good DSSG Project?*].

1. *A solvable problem.* This project cannot solve poverty, but it can help alleviate it by reducing corruption, collusion and fraud.
2. *A challenging problem.* Challenging problems encourage teamwork, spawn creative solutions, and play a key role in a Data Scientist's ability to solve real-world problems and an understanding, excitement, and passion for solving problems with social impact. For example, this World Bank project involved search-engine expertise to find links between corrupt applicants online.
3. *An important problem with social*

*impact.* Working in a big project is an investment, not only financially, but also opportunistically (when someone choose to do a project, someone is choosing not to do another). It is important to dedicate the limited resources to substantial problems. Each project must meet an operational need for the partner organization and must have a tangible connection to “social good.” For example, this project helps more people over fewer people and that solve chronic problems over temporary problems.

4. *A motivated, capable, and committed partner.* No project can succeed without a fully invested project partner. Project partners understand the problem, they have subject-matter expertise, and they ultimately decide how a Data Scientist’s work is used. Partners often look at the problem differently, which is important for solving tough problems. Partners provide insight into the problem and to guide Data Scientist to develop a solution. This demands a lot from partners. It often requires partners stretching themselves and asking themselves hard questions. It also requires time. For this project, the World Bank helped scope the problem before it started, they gave a complete presentation about their work, they were available for at least once a week for feedback and discussion during the duration of the project, and they are actually using the results of this work when it finished.

5. *Appropriate, relevant data.* Getting the data a project needs is almost always the biggest challenge. Important things go unmeasured or unrecorded or, more commonly, cannot be shared. Many projects involve sensitive information. Getting lawyers to agree on data and code sharing can take months. It is important for Data Scientist to be flexible. Partners have anonymized data (while keeping it useful at an individual level), conducted background checks, and require Data Scientist to do analyses on their internal computer systems (remotely). All this while maintaining a spirit of openness. In this case, the World Bank contributed with all the relevant data they have in order to build a solution that’s appropriate, effective, and easily deployed.

## 1.1 The World Bank

The World Bank was created in 1944 as one institution and has evolved into an association of five development institutions. It is composed by the International Bank for Reconstruction and Development (IBRD), the International Development Association (IDA), the International Finance Corporation (IFC), the Multilateral Guarantee Agency (MIGA), and the International Center for the Settlement of Investment Disputes (ICSID). In its interior it was formed by a large group of engineers and financial analysts who work from Washington, D.C. Now, they have a very heterogeneous and distinct staff that includes economist, public policy experts, sector experts and social scientists and about one third of the workers has spread to the country offices. They have more than 10,000 employees in more than 120 offices worldwide. While reconstruction remains an important part of their objectives, however, at today's World Bank, poverty reduction through an inclusive and sustainable globalization remains the overarching goal. For details see [web page *History of the World Bank*].

The World Bank Group has set two goals for the world to achieve by 2030; first, end extreme poverty by decreasing the percentage of people living on less than \$1.90 a day to no more than 3%; second, promote shared prosperity by fostering the income growth of the bottom 40% for every country; third, the World Bank is a vital source of financial and technical assistance to developing countries around the world. The World Bank Group is not a bank in the ordinary sense but a unique partnership to reduce poverty and support development [web page *What we do*].

The World Bank Group provides Financial Products and Services by giving out low-interest loans, zero to low-interest credits, and grants to developing countries.

“These support a wide array of investments in such areas as education, health, public administration, infrastructure, financial and private sector development, agriculture, and environmental and natural resource management. Some of our projects are cofinanced with governments, other multilateral institutions, commercial banks, export credit agencies, and



private sector investors.” [web page *What we do*].

They offer financing through trust fund partnerships with bilateral and multilateral donors. For example, many partners have asked the Bank to help manage initiatives that address needs across a wide range of sectors and developing regions. Projects vary widely in scale and scope, ranging from developing hydropower systems<sup>1</sup> to rehabilitating coral reefs<sup>2</sup> to improving roads, health, education and agriculture systems<sup>3</sup>.

This work focuses in these Financial Products and Services. It analyzes their contracts (see the chapter 2 for more details) by using historical data on over 300,000 major contracts funded by World Bank in the form of low-interest loans, zero to low-interest credits, and grants to developing countries covering from the past 20 years. Historical data including such features as company name, country, sector, and total award amount.

The World Bank also provides Innovative Knowledge Sharing, by offering support to developing countries through policy advice, research and analysis, and technical

<sup>1</sup> WASHINGTON, March 20, 2014-Sub-Saharan Africa is blessed with large hydropower resources that can bring electricity to homes, power businesses and industry, light clinics and schools, and spur economic activity, creating jobs and improving human well-being. Yet, only 10% of this hydropower potential has been mobilized, weakening the fight to end poverty and boost shared prosperity on the continent. See web page *Transformational Hydropower Development Project Paves the Way for 9 Million People in the Democratic Republic of Congo to Gain Access to Electricity*

<sup>2</sup> Wakatobi, Indonesia, June 5, 2014 - As the world’s largest archipelago, Indonesia is blessed with at least 5.1 million hectares of coral reefs. However, almost 65 percent of the reefs are now considered threatened from overfishing. Almost half are considered threatened specifically from destructive fishing practices. Nadjib Prasyad runs the Fisheries Office in Wakatobi, South-east Sulawesi, and he laments the various activities that destroy the reefs and consequently threaten the livelihood of the villages: fish bombing, sand extraction, collection of the reefs themselves. Prasyad says that, once the reefs die, so do the fish: “We have nothing except our coral reefs. So we have to really protect them since they’re the only source of our region’s development. See web page *It Takes Villages to Conserve Indonesia’s Precious Coral Reefs*

<sup>3</sup> “We didn’t use to have bus service to the communities; now we have it every hour.” “Now we have a paved highway and signs; it’s comfortable for traveling.” “When there was a drought, there was no water for the animals, for the pasture, for irrigating the produce we consumed and much less for crops for sale.” “With the water, people returned to the community and it is improving their quality of life.” Today, these are some of the phrases that can be heard on the roads of Chimborazo, one of the largest provinces in Ecuador’s central highlands. See web page *New roads and irrigation systems improve life in Ecuador*

assistance.

“The analytical work often underpins World Bank financing and helps inform developing countries’ own investments. In addition, we support capacity development in the countries we serve. We also sponsor, host, or participate in many conferences and forums on issues of development, often in collaboration with partners” [web page *What we do*].

This is why this project becomes of high value for the World Bank Group. On one hand, this work helps them to provide much better Financial Products and Services by reducing the cost that corruption, collusion and fraud generates in the process of lending money to developing countries throughout the world’s developing countries. On the other, this project contribute to their Innovative Knowledge Sharing by supplying them with the necessary tools and algorithms they need in order to share corruption, collusion and fraud analysis’ results.

This project relies on the fact that the World Bank group, in order to ensure that countries can access the best global expertise and help generate cutting-edge knowledge, is constantly seeking to improve the way it shares its knowledge and engages with clients and the public at large. This includes measurable results, improving every aspect of their work like how projects are designed, how information is made available, and how to bring our operations closer to client governments and communities and also includes their Open Development, a set of free, easy-to-access tools, research and knowledge to help people address the world’s development challenges. This work would not be possible without all this previous work and is a virtuous circle because it contributes by helping them achieve this goal.

This work uses the Open Data<sup>4</sup> website from the World Bank, which offers free access to comprehensive, downloadable indicators about development in countries around the globe, data needed in order to detect and predict patterns of collusion, corruption, and fraud in its contracts data by using models of contract-specific risk.

On chapter 2 BLA BLA BLA

---

<sup>4</sup>Go to web page *Data Bank*

On chapter 2

---

## **Chapter 2.**

# **The World Bank's Procurements**

---

As shown in the chapter 1, The World Bank Group provides Financial Products and Services by giving out low-interest loans, zero to low-interest credits, and grants to developing countries. They offer financing through trust fund partnerships with bilateral and multilateral donors. This work focuses in these Financial Products and Services. It analyzes their contracts by using historical data on over 300,000 major contracts funded by World Bank.

In this chapter BLA BLA BLA...BLA BLA

## **2.1 The Wold Bank Group's Rules and Guidelines**

The process by which the World Bank Group gives out low-interest loans, zero to low-interest credits, and grants to developing countries is not simple. They have a set of guidelines, available online [The World Bank, 2011a], in which they specify how to apply to all the contracts for goods, works and services financed in whole or in part from the Bank's loans, credits and grants which would thereby cover both International Bank for Reconstruction and Development (IBRD) and the International Development Association (IDA). Similar provisions apply for the selection of

Consultants under the Consultants Guidelines [The World Bank, 2011b]. For the purpose of this project the Consultants will not be deeply explained.

In the World Bank's rules and guidelines they consider "goods" and "works" all the related services such as transportation, insurance, installation, commissioning, training, and initial maintenance. Also, "goods" includes commodities, raw material, machinery, equipment, vehicles, and industrial plant.

"This requirement is made applicable in each operation through the loan, credit or grant agreement thus making their use a legal obligation on the part of the Borrower. It applies as well to counterpart funds that are also used to finance contracts with loan/credit proceeds. In addition, parts of Bank-financed projects that are co-financed by other donors are also subject to Bank rules. This could also be the case under parallel financing when co-financiers agree to using Bank guidelines (leverage effect)." [web page *Procurements Rules*]

In order to ensure that these legal provisions are observed during project implementation, Bank staff review and provide a "no objection" before the Borrower implements certain procurement decisions<sup>1</sup>. Bank staff will review all the contracts that have not been subject to Prior Review so as to ensure that the Guidelines were applied as required. Where this is found not to be the case, the Bank applies various remedies including *mis-procurement*, which in most instances requires cancellation and refund of the funds in question<sup>2</sup>.

### 2.1.1 Procurements General Considerations

In every Procurement the World Bank expect that the responsibility for the implementation of the project, and therefore for the award and administration of contracts

---

<sup>1</sup>For all International Competitive Bidding (ICB) and other high-value, high-risk, and otherwise complex contracts

<sup>2</sup>Advertising requirements vary depending on whether Procurement is subject to International Competitive Bidding (ICB) or National Competitive Bidding(NCB).

under the project, rests with the Borrower. The Bank, is required ensure that the proceeds of any loan are used only for the purposes for which the loan was granted, with due attention to considerations of economy and efficiency and without regard to political or other non-economic influences or considerations. While in practice the specific procurement rules and procedures to be followed in the implementation of a project depend on the circumstances of the particular case, four considerations generally guide the Bank's requirements:

- a) "the need for economy and efficiency in the implementation of the project, including the procurement of the goods, works, and non-consulting services involved;
- b) the Bank's interest in giving all eligible bidders from developed and developing countries the same information and equal opportunity to compete in providing goods, works, and non-consulting services financed by the Bank;
- c) the Bank's interest in encouraging the development of domestic contracting and manufacturing industries in the Borrowing country; and
- d) the importance of transparency in the procurement process."<sup>3</sup>

### **2.1.2 Procurement Plan**

Any firm or person interested in a procurement has to submit a Procurement plan to the World Bank in which they specify general characteristics of the project such as the Bank's approval date of the procurement plan, the date of general procurement notice and the period covered by the procurement plan. It has to include the details concerning the goods and works and non-consulting services required for the project such as the Prior Review Threshold, prequalification, proposed procedures for CDD components, reference to (if any) project operational/procurement manual, any other special procurement arrangements, and a summary of the procurement packages planned during the first 18 months after project effectiveness. With regard to the

---

<sup>3</sup>See The World Bank, 2011a.

consultants, it has to include a prior review threshold, short list comprising entirely of national consultants, any other special selection Arrangements (including advance procurement and retroactive financing, if applicable or delete if not applicable), and consultancy assignments with selection methods and time Schedule. The appendix A shows a sample procurement plan to be filled [web page *Procurement Plan Template*].

The World Bank's guidelines for procurement of goods, works, and non-consulting services [The World Bank, 2011a] establishes that open competition is the basis for efficient public procurement. In their articles, they state that each Borrower shall select the most appropriate method for the specific procurement. They have to choose between International Competitive Bidding (ICB) or National Competitive Bidding (NCB). This means, in most cases, ICB, properly administered, and with the allowance for preferences for domestically manufactured goods and, where appropriate, for domestic contractors for works under prescribed conditions is the most appropriate method. The Bank requires its Borrowers to obtain goods, works, and non-consulting services through ICB open to eligible suppliers, service providers, and contractors<sup>4</sup>.

In order to satisfy these requirements, any procurement needs to be transparent, competitive and legal according to the guidelines the World Bank established. Contractors providing goods and services on World Bank projects are typically hired through a competitive bidding process. The problem is that in practice, the Borrowers and Contractors tend not to fulfill those requirements. Occasionally, prospective contractors influence the competitive system by colluding with other contractors, bribing government officials, or otherwise manipulating the bidding process. These offenses have far-reaching effects on the price and quality of contract delivery. There is a considerable amount of cases in which corruption, collusion and/or fraud takes place. The guidelines include some sections to address those problems. The World Bank is committed to detecting instances of collusion, corruption, and fraud in order to maximize its global impact.

---

<sup>4</sup>Section II of these Guidelines describes the procedures for ICB [The World Bank, 2011a].

## 2.2 Corruption, Collusion and Fraud

The World Bank's guidelines for procurement of goods, works, and non-consulting services [The World Bank, 2011a] establishes what's to be understood by corruption, collusion, fraud, coercion and obstruction.

It is the Bank's policy to require that Borrowers<sup>5</sup>, bidders, suppliers, contractors and their agents<sup>6</sup>, sub-contractors, sub-consultants, service providers or suppliers, and any personnel thereof, observe the highest standard of ethics during the procurement and execution of Bank-financed contracts. In pursuance of this policy, the Bank defines, the terms set forth below as follows<sup>7</sup>:

**Definition 2.2.1.** A *corrupt practice* is the offering, giving, receiving, or soliciting, directly or indirectly, of anything of value to influence improperly the actions of another party.

**Definition 2.2.2.** A *collusive practice* is an arrangement between two or more parties designed to achieve an improper purpose, including to influence improperly the actions of another party.

**Definition 2.2.3.** A *fraudulent practice* is any act or omission, including a misrepresentation, that knowingly or recklessly misleads, or attempts to mislead, a party to obtain a financial or other benefit or to avoid an obligation.

**Definition 2.2.4.** A *coercive practice* is impairing or harming, or threatening to impair or harm, directly or indirectly, any party or the property of the party to influence improperly the actions of a party.

**Definition 2.2.5.** A *obstructive practice* is deliberately destroying, falsifying, altering, or concealing of evidence material to the investigation or making false statements to investigators in order to materially impede a Bank investigation into allegations

---

<sup>5</sup>including beneficiaries of Bank loans.

<sup>6</sup>whether declared or not.

<sup>7</sup>Definitions taken from The World Bank, 2011a. Updated to 2011.



of a corrupt, fraudulent, coercive or collusive practice; and/or threatening, harassing or intimidating any party to prevent it from disclosing its knowledge of matters relevant to the investigation or from pursuing the investigation, or acts intended to materially impede the exercise of the Bank's inspection and audit rights.

In its effort to avoid practices from definitions 2.2.2 to 2.2.5, the World Bank will respond accordingly to each contract in question. The Bank will *reject a proposal* for award if it determines that the bidder recommended for award, or any of its personnel has, directly or indirectly, engaged in corrupt, fraudulent, collusive, coercive, or obstructive practices in competing for the contract in question. The Bank will *declare mis-procurement* and cancel the loan if it determines at any time that representatives of the Borrower engaged in any corrupt, fraudulent, collusive, coercive, or obstructive practices during the procurement or the implementation of the contract. The Bank will *sanction* a firm or individual, including by publicly declaring such firm or individual ineligible, either indefinitely or for a stated period of time to be awarded a Bank-financed contract; and to be a nominated sub-contractor of an otherwise eligible firm being awarded a Bank-financed contract. The Bank will require that a clause be included to permit the Bank to inspect all accounts, records, and other documents relating to the submission of bids and contract performance, and to have them audited by auditors appointed by the Bank.

### **2.2.1 Examples: What is Fraud and Corruption?**

An example of fraud according to the definition 2.2.3:

“In a few years since its establishment, a consulting company is awarded multiple World Bank Group-financed contracts totaling in millions of dollars. The proposals submitted by the consulting company contain numerous past project experiences and references that contribute to its success and eligibility. However, a review of the consulting company's past project experiences reveal that the company has claimed the experi-

ences of individual consultants as its own, as well as grossly exaggerated the value of the projects that it has undertaken. Not only is the consulting company misrepresenting its qualifications and capacity, it is also cutting its subconsultants' contracts by half, but claiming the full amount on its invoices to the client. The quality of the deliverables from this consulting company is highly questionable, which affects the project's overall developmental goals." [web page *What is Fraud and Corruption?*]

An example of corruption according to the definition 2.2.1:

"A supplier agrees to pay kickbacks to a senior government official through an agent it hires as a sub-consultant to perform business development and marketing services but without any deliverables. This agent is connected to a senior government official who is demanding a commission from every bidder as the official has influence over the bid evaluation committee and can steer the award of the contract to any bidder willing to pay. This supplier builds in the kickback amount as a percentage of the contract value, and pays for it from the funds it receives from the World Bank Group-financed project. Project financing costs are artificially inflated by these practices, and the supplier recovers costs by providing less expensive and lower quality goods." [web page *What is Fraud and Corruption?*]

An example of collusion according to the definition 2.2.2:

"A project official arranges to steer contracts on a World Bank Group-financed project to his own company and that of his relatives. The project official not only tells his relatives' companies what prices to put in their bids, but also what particular technical specification to include. The bids of companies that are not part of this inner circle are disqualified as being technically non-responsive, leaving the project official's and his relatives' companies as the lowest evaluated bidders on the different contracts. Not only is the integrity of the procurement process compromised,

but the winning bid prices are considerably higher than what would have been with genuine competitive bidding.” [web page *What is Fraud and Corruption?*]

An example of coercion according to the definition 2.2.4:

“A contractor is stopped from submitting his bid at the bid opening. Persons connected to a competitor block the contractor from entering the building where the bid opening is taking place, and tell him that ‘if he cares for his family, he should not submit a bid.’ Another bidder who comes to submit a bid is also stopped by these same persons who tells the bidder that ‘it is not his turn to win this contract.’ The two bidders leave the bid opening and do not submit a bid out of fear.” [web page *What is Fraud and Corruption?*]

An example of coercion according to the definition 2.2.4:

“World Bank Group investigators contact a company alleged to have paid a bribe on a World Bank Group-financed contract and request to audit the company’s financial records. The company refuses to do so despite its agreement and obligation under the contract to allow the World Bank Group access to these records. Furthermore, it withholds key documents and alters other documents that are given to the investigators.” [web page *What is Fraud and Corruption?*]

## 2.3 World Bank Listing of Ineligible Contractors

The World Bank has a long history of giving loans for procurement of goods, works, and non-consulting services based on the four considerations listed above. As expected, there has been a long history of malicious activity in the procurement process. This is why The World Bank publishes a list of contractors which had been discovered to have a practice that lead to either a rejection, a mis-procurement, an

obstructive practice or a sanction. This project refers to this list as the debarment data.

The World Bank publishes an online [web page *World Bank Listing of Ineligible Firms and Individuals*] list that includes all the firms and individuals listed as ineligible to be awarded a World Bank-financed contract for the periods indicated because they have been sanctioned under the Bank's fraud and corruption policy as set in the Procurement Guidelines or the Consultant Guidelines. Such sanction was imposed as the result of an administrative process conducted by the Bank that permitted the accused firms and individuals to respond to the allegations<sup>8</sup>, or a cross-debarment<sup>9</sup> made effective by the World Bank, Asian Development Bank, European Bank for Reconstruction and Development, Inter-American Development Bank, and African Development Bank.

The table 2.1 bellow shows a short sample of how is the debarments list published. As it can be seen from the table 2.1, it includes a *Firm Name*, *Address*, *Country*, *Ineligibility Period* and *Grounds*. The *Firm Name* can be a person or an entity, the *Ineligibility period* corresponds to the time frame in which that specific *Firm Name* is debarred from receiving any type of award financed by the World Bank, and finally, the *Grounds* refers to the violations made to the Consultants Guidelines (CG) or the Procurements Guidelines (PG) and their specific article and section.

---

<sup>8</sup>“Through July 2007, this process was conducted in accordance with the Sanctions Committee Procedures adopted on August 2, 2001. Since then, the process has been conducted in accordance with the Sanctions Procedures of the World Bank Group Sanctions Board.”

<sup>9</sup> Since 2011.

Table 2.1. Debarred Firms and Individuals (sample)

Firm Name	Address	Country	Ineligibility Period		Grounds
			From	To	
MR. NGUYEN PHUONG QUY*298	TOWER CENTER OFFICE BUILDING, NO. 83A LY THUONG KIET STREET, DISTRICT HOAN KIEM, HANOI	Vietnam	15-DEC-2015	14-DEC-2026	CG, 1.22(a)(ii); PG, 1.14(a)(ii)-(iii)
SFC VIETNAM INVESTMENT DEVELOPMENT FOR ENVIRONMENT CORPORATION*297	TOWER CENTER OFFICE BUILDING, NO. 83A LY THUONG KIET STREET, DISTRICT HOAN KIEM, HANOI	Vietnam	15-DEC-2015	14-DEC-2025	CG, 1.22(a)(ii); PG, 1.14(a)(ii)-(iii)
MR. NIKOLAI GEORGIEVITCH OBRADOVITCH*296	ANTONIENKO 3,-43, 190000 SAINT PETERSBURG	Russian Federation	30-OCT-2015	29-OCT-2019	CG, 1.22(a)(i)

**Source:** see the web page *World Bank Listing of Ineligible Firms and Individuals*.

**Notes:** Several firms above are marked with an asterisk (\*). The period of ineligibility of the sanctioned firm extends to any firm directly or indirectly controlled by the sanctioned firm.

The World Bank also publishes a second list that includes the *Other Sanctions* that includes firms that are typically under a *Conditional Non-debarment*. “The Conditional Non-debarment means that so long as the sanctioned entity meets certain conditions, including (a) implementing a corporate compliance program acceptable to the Bank; (b) fully cooperating with the Bank; and (c) not attempting to evade the sanction imposed on the firm and those entities it directly or indirectly controls, the sanctioned entity will continue to be eligible to participate in Bank-financed activities.” [web page *World Bank Listing of Ineligible Firms and Individuals*]; or has a *Letter of Reprimand* by which the World Bank officially notifies the firm its sanction.

## 2.4 Investigations

The World Bank conducts some investigations to determine whether to sanction or not an entity within the procurement process. The investigations are primarily based upon the allegations they receive, so it is extremely important that those people who are involved in activities supported by World Bank Group funds take the initiative to report suspected fraud or corruption. By now, the World Bank has a basic way of conducting investigations regarding corruption, collusion, fraud, coercion and obstruction. The web page *Investigations* lists three types of investigations by which the Bank analyze the sanctionable practices.

1. *Complaint Intake*, where Integrity Vice Presidency (INT) performs an initial assessment of every complaint that it receives. This assessment determines whether: the complaint relates to a sanctionable practice in World Bank Group-financed projects, the complaint has credibility and the matter is of sufficient gravity to warrant an investigation. Complaints outside of INT's jurisdiction are redirected to other areas of the World Bank Group as appropriate. Complaints that fall under INT's jurisdiction are investigated if they are determined to be of a higher priority. When a complaint does not reach this threshold, INT works with Operational staff to address the issues raised. In assigning priority, INT also considers the possible reputational risk to the World Bank Group, amount of funds involved and quality of the information or evidence in INT's possession.
2. *Investigation of Cases*, where, through investigations, INT ascertains whether firms and/or individuals have engaged in one of the World Bank Group's five sanctionable practices. Since an INT investigation is administrative in nature, the standard of proof is akin to a "balance of probabilities" and therefore lower than the criminal standard of "beyond a reasonable doubt." The World Bank Group, for that reason, has to prove that it is more likely than not that the alleged misconduct has occurred. If INT finds sufficient evidence to prove the allegation, the allegation is considered substantiated. The allegation is considered

unsubstantiated if there was insufficient evidence to prove or disprove it, and unfounded if the allegation has no basis in fact.

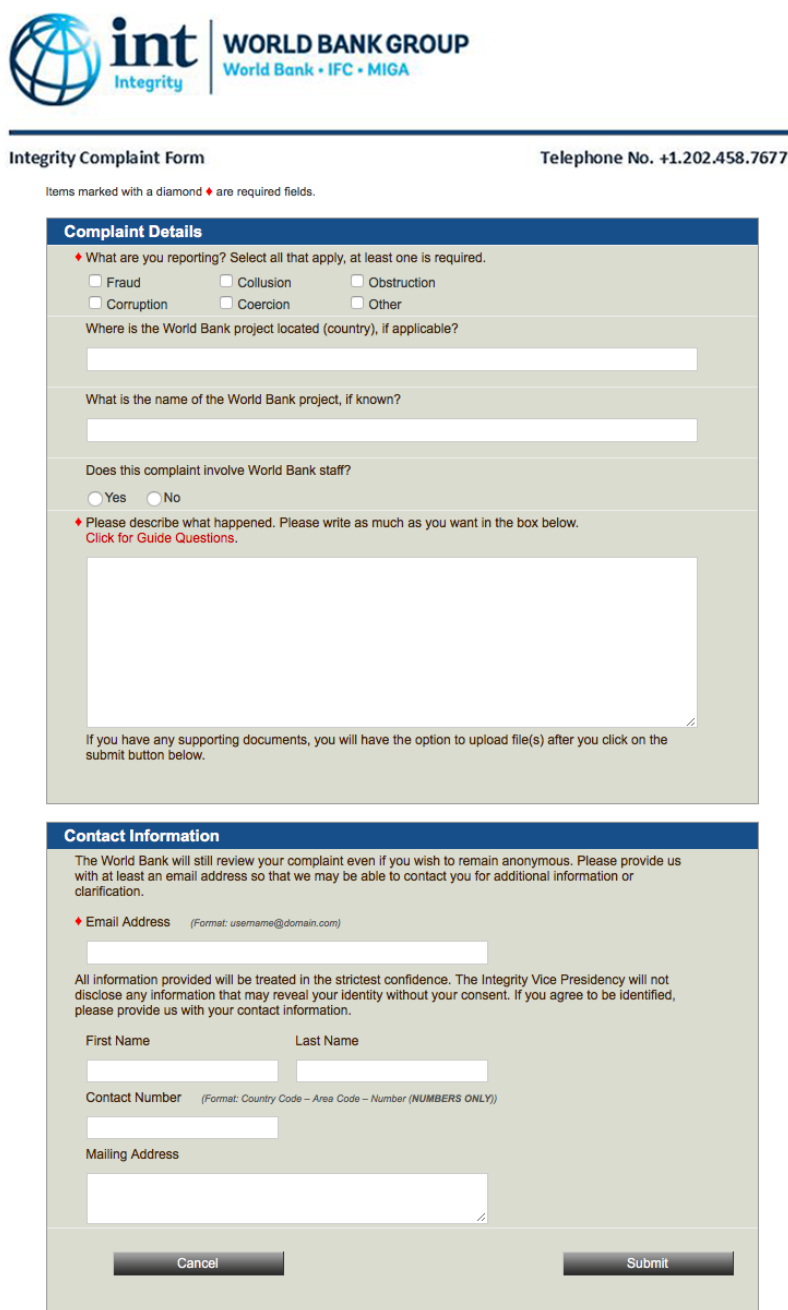
3. *Investigation Reports*, when INT substantiates a case, it produces a Final Investigation Report (FIR). In some cases, INT will produce an FIR even if there is not reasonably sufficient evidence to substantiate a complaint - for example, if INT believes that the investigation unearthed important lessons that should be shared with colleagues in the World Bank Group. FIRs are sent to regional management for comment before being finalized and provided to the World Bank Group President. INT strives to ensure that the maximum time between opening a case and completing an investigation report is twelve months for normal cases and eighteen months for complex cases. FIRs also form the basis for two other INT outputs: referral reports, which INT sends to relevant national authorities if evidence indicates that the laws of a World Bank Group member country may have been violated; and redacted reports, which are provided to the World Bank Group's Board of Executive Directors and, after the completion of any related sanctions proceedings, posted on this site.

### **2.4.1 Report Suspected Fraud or Corruption**

The World Bank's Integrity Vice Presidency (INT) is in charge of taking care of any allegation that involves possible fraud, corruption, collusion, coercion and obstruction in World Bank-funded projects and/or against World Bank staff. To Report an allegation, the World Bank provides two main options, submit an Online Integrity Complaint Form or download the free Integrity Application.

The Integrity Complaint Form is a way to send a report to INT. The figure 2.1 bellow shows the format of an Integrity Complaint Form. As the figure shows, it is a secure and confidential third-party website that is managed by INT. The form requires information regarding your allegations and a summary of the concerns. There's an option to upload supporting documents after submitting the Form.

Figure 2.1. Integrity Complaint Form



**int** | **WORLD BANK GROUP**  
Integrity | World Bank • IFC • MIGA

---

**Integrity Complaint Form** Telephone No. +1.202.458.7677

Items marked with a diamond ♦ are required fields.

**Complaint Details**

♦ What are you reporting? Select all that apply, at least one is required.

<input type="checkbox"/> Fraud	<input type="checkbox"/> Collusion	<input type="checkbox"/> Obstruction
<input type="checkbox"/> Corruption	<input type="checkbox"/> Coercion	<input type="checkbox"/> Other

Where is the World Bank project located (country), if applicable?

What is the name of the World Bank project, if known?

Does this complaint involve World Bank staff?

☐ Yes ☐ No

♦ Please describe what happened. Please write as much as you want in the box below.  
[Click for Guide Questions.](#)

If you have any supporting documents, you will have the option to upload file(s) after you click on the submit button below.

**Contact Information**

The World Bank will still review your complaint even if you wish to remain anonymous. Please provide us with at least an email address so that we may be able to contact you for additional information or clarification.

♦ Email Address (Format: username@domain.com)

All information provided will be treated in the strictest confidence. The Integrity Vice Presidency will not disclose any information that may reveal your identity without your consent. If you agree to be identified, please provide us with your contact information.

First Name  Last Name

Contact Number (Format: Country Code – Area Code – Number (NUMBERS ONLY))

Mailing Address

**Source:** web page *Integrity Complaint Form*



## 2.5 Searching for *red-flags*

The real issue for the World Bank becomes that, in order for them to conduct a Complaint Intake, an Investigation Case, or even an Investigation Report, they rely on third parties. As seen in online form that the World Bank provides, figure 2.1 and on the mobile application, they are making a significant effort to supply third parties with the best tools to report any cases of corruption, fraud, collusion, coercion and obstruction so that reporting do not becomes the obstacle. Said that, the real problem for The World Bank becomes that their efforts to conduct investigations is limited to the information they receive from third parties. This is an important matter when their objective is to reduce the world poverty by awarding procurements, loans and credit to developing countries and at the same time they are facing corruption, fraud and collusion problems by which only few sectors are being benefited.

Investigators at the World Bank's Integrity Vice Presidency (INT) are responsible for investigating collusion, as well as the other sanctionable offenses of coercion, fraud and corruption in World Bank-financed projects. The challenge with a global portfolio is how to detect collusion or corruption, particularly if no one reports the incident.

For this reason, The World Banks has made an effort in order to conduct a new type of investigations that do not depend on third parties but depends on data. The problem now becomes: how to identify corruption, collusion, fraud, coercion and obstruction by looking on data? The World Bank has some initial thoughts and has been trying to search for new and innovative ways to tackle this problem. For example, on June 2014 they organized a DataSprint event that helped them to dive into data to address these type problems. The objective was to analyze, visualize, and mash-up the the data. I went to Washington, D.C. for several days of meetings and collaboration with the project partners at INT and the World Bank's Office of the Controller. The trip was an opportunity to learn what corruption, fraud, and

collusion looks like in the data.

INT investigators have identified several patterns in bidding behavior that could point to collusion or other forms of corruption. Regular, periodic rotation of contract awards among a small group of contractors, for example, might indicate that bidders are working together to set prices or distribute contracts. Contractors may split large projects into multiple small ones, keeping individual contract values below a threshold that triggers additional review or competitive bidding. The total number of bidders shrinking over time might indicate that potential contractors are being pushed out of the market through some form of illegal coercion. While none of these findings represent clear-cut evidence of collusion, they are detectable clues that can help direct investigative efforts and ensure that the integrity of development is not compromised [web page *Clean Development: Data Mining for Corruption Risks*].

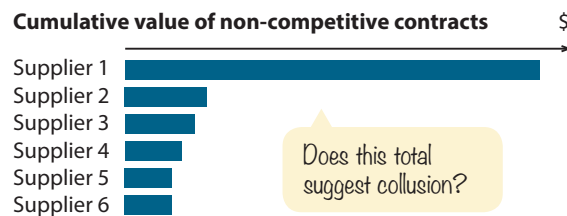
The INT developed a visualization, shown on figure 2.2, that show cases that suggest malicious practices. To start with, sub-figure 2.2a shows data that corresponds to a case in which six suppliers that are supposed to be competitive and all the cumulative value on non-competitive contracts are being assigned to supplier number one. On sub-figure 2.2b it is shown that contracts, that correspond to Company A, numbers one and two overcome the threshold for a common competitive bidding. This might be suggesting collusion. Sub-figure 2.2c shows a case where suppliers four to eight always bid a little less than the winning bid suggesting another type of collusion. Now, on sub-figure 2.2d there is a suggestion of favoritism that suggest corruption or even coercion in which the value of a non-competitive contract goes high exactly in the times of a governmental election. Sub-figure 2.2e suggests that companies that have contracts or directors in common with other companies that have been previously debarred are suspicious of malicious activities. As it can be seen from sub-figures 2.2f and 2.2h there tend to be bidding patterns in which take turns by bidding differently in each round such that they take turns to get the contracts. This affects directly the process making it non competitive. Lastly, it can be seen on figure 2.2g that as time pass, only two suppliers are the ones participating on

proposals among a vast pool. This fact is suggesting that those two suppliers might have done something to coerce the others not to participate on proposals after time four.

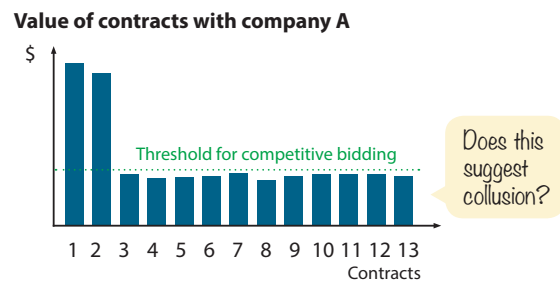
Figure 2.2. What does corruption look like?

## What does **corruption** look like?

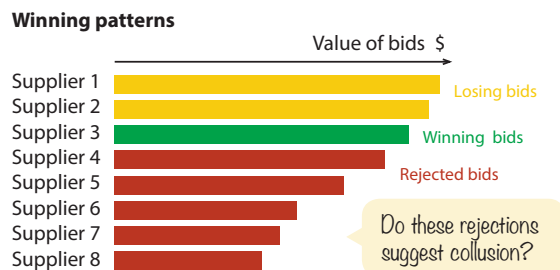
(a) Cumulative value of non-competitive contracts



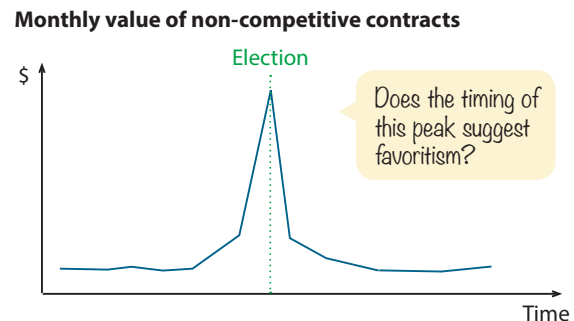
(b) Value of Contracts



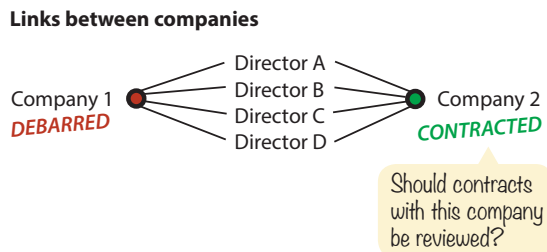
(c) Winning patterns



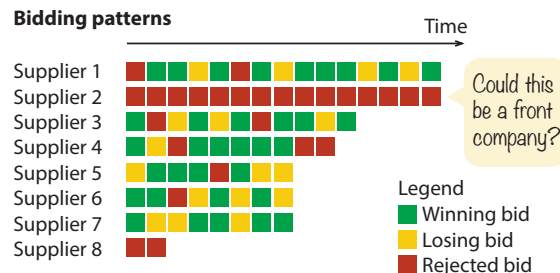
(d) Monthly value of non-competitive contracts



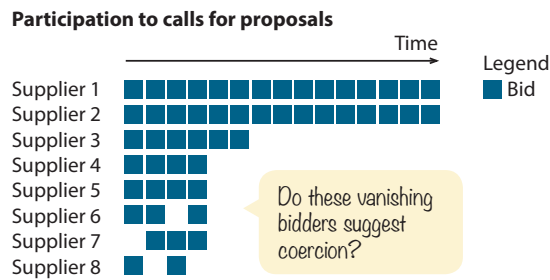
(e) Links between companies



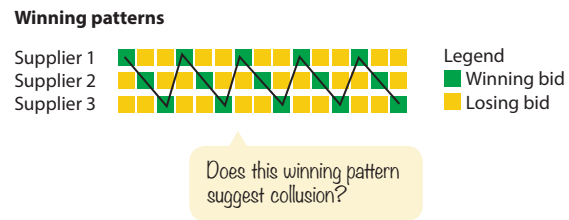
(f) Bidding patterns



(g) Participation to call for proposals



(h) Bidding patterns



**Source:** Figures taken from Gagnon and Wiramidjaja, 2014

Another priority during the trip to Washington D.C. was asking questions and learning about the data that will provide a window into the problems that INT is working to address. To data scientists, diving into analysis and model-building right away can be an enticing prospect when presented with an exciting new data set, but engaging with stakeholders and domain experts at the earliest stages can help guidelinesde analysis and save considerable time down the road.

---

## Chapter 3.

# The World Bank's Data

---

As shown on chapter 1, collecting the necessary data for any Data Science project is almost always a huge task because the reality tends to be that some of the most relevant variables commonly go unmeasured or are private. This project uses two types of data, private and public. This chapter describes the two main types of data used for this project. Section 3.1.1 describes the characteristics of private data and the scope it has. It also explains about the confidentiality agreement Integrity Vice Presidency (INT) offered this project in exchange for investigations data. Section 3.2 explains how to download public data from the World Bank's servers using coding tools.

## 3.1 Private data

### 3.1.1 Investigations

The most valuable data for this project was the private data which the partners at INT made available for the project while maintaining a spirit of openness. A confidentiality agreement was signed in order to give access to the investigations

database. The only condition was that all the data and analyses had to be stored and processed on a remote server with an encryption and security measurements that satisfied the World Bank’s standards.

In this case the server was an Amazon Web Service’s (AWS), Amazon Elastic Compute Cloud (Amazon EC2) machine. EC2 is a web service that provides resizable compute capacity in the cloud. It is designed to make web scale cloud computing easier for developers<sup>1</sup>. For more details on how to setup an AWS computer see [Amazon Web Service, 2015] and appendix B. In other words, the data was stored in an AWS machine on the cloud that had an encrypted folder with the specific files. Since all the data had to stay on the cloud, all the processing had to be done on the cloud too. For this purpose, a server running `Python` and `R` was hosted at an AWS computer.

## Description

The investigations data consists of nearly 13,000 cases, of which over half are Fraud and Corruption. 377 are labeled as Collusion. These are from the “Allegation Category” column, Notably, allegation categories include Fraud, Corruption, and Fraud & Corruption. It is possible for a case’s description to just say Fraud and for its allegation category to still be Fraud and Corruption. For the purpose of this project, each of these 3 categories were considered as the same (i.e. all “Fraud and Corruption”). Allegations also include a “Allegation Sub-Category”. Collusion is sometimes listed as a sub-category, for example: Allegation Category: Fraud and Corruption or Allegation Sub-Category: Bid Manipulation/Collusion. The Unit of analysis is the investigations data is the “Subject” which refers to the entity that was investigated.

---

<sup>1</sup>“Amazon EC2’s simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon’s proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use. Amazon EC2 provides developers the tools to build failure resilient applications and isolate themselves from common failure scenarios.” [web page *Amazon EC2 - Virtual Server Hosting*]

Mostly this variable refers to companies and people, but on occasion other entities on a project may be subjects of investigation, for example the coordinating agency. For example Case # A-AAA-1234-5678, the subject is: Country's X Ministry of Education. In procurement data it shows up in "Implementing Agency" differently.

The Labels in the investigations are very different and specific to each investigation case. The most relevant label will be the Allegation Outcome's column. The most frequent outcomes for a case are "Substantiated", which means that the Subject was found to be guilty of the Allegation, "Unfounded", which means that the Subject was found to be innocent of the allegation, and "Unsubstantiated", which means that the investigation did not have the necessary means to conclude that the subject was guilty, but neither does it have to conclude that the subject is innocent. Substantiated and Unsubstantiated occur in roughly equal quantities, while Unfounded is fairly rare. There are other labels like "We kicked it to another organization".

The identities of the cases within the investigations data include that about 80% of the data has either the name of the accused (the majority) or the project number (about a third). These are in the Subject and Project Number columns, respectively. That means about 20% of the data appears to be "lost", in that we can't match it up to any procurement data. There is potentially other information in the "Title" field, but it is unstructured. There could theoretically be information in other fields, but they are not as rich.

There are among 1000 different sectors but most of them are mixed sectors and have not been previously cleaned. For example, when an investigation involves cases that cover sectors such as transportation and urban development there can be the case that you find a Transportation sector, a Urban Development sector and a Transportation & Urban Development Sector. For the specifics of the model, this project works with sector specific as they are, this meaning that, for the project, the three cases mentioned before are considered to be different sectors. By doing so, there are about 205 sectors.

## 3.2 Public data

Fortunately for the project, not all the data was private. This work uses the Open Data<sup>2</sup> and the Data Bank websites from the World Bank, which offer free access to comprehensive, downloadable indicators about development in countries around the globe.

Public data from the World bank comes from different databases. They are generated at different sections within the World Bank so each indicator can be obtained from a specific database according to its nature. For example, all the Development indicators can be obtained at the World Bank's World Development Indicators database, indicators regarding how easy it is to do business in each specific country could be obtained from the Doing Business database and indications regarding world's populations can be obtained at the Health Nutrition and Population Statistics database.<sup>3</sup>

A good practice in Data Science is to generate code so that all results can be easily replicated. That was the reason why, for this project all the data that the World Bank bank provides publicly by an Application Programming Interface (API) was collected with reliable code. Thanks to prior World Bank's work, accessing the World Bank Data APIs with code in languages such as **Python**, **R**, **Ruby** and **Stata**

---

<sup>2</sup>Go to web page *Data Bank*.

<sup>3</sup>Different databases include: Africa Development Indicators, Statistical Capacity Indicators, Country Policy and Institutional Assessment (CPIA), Country Partnership Strategy for India, Corporate Scorecard, Doing Business, Exporter Dynamics Database: Country-Year, Education Statistics, Enterprise Surveys, Global Findex ( Global Financial Inclusion database), G20 Basic Set of Financial Inclusion Indicators, Gender Statistics, Global Economic Monitor, GEP Economic Prospects, Global Financial Development, Global Economic Monitor (GEM) Commodities, Global Partnership for Education, Global Social Protection, Health Nutrition and Population Statistics, Health Nutrition and Population Statistics by Wealth Quintile, Health Nutrition and Population Statistics: Population estimates and projections, International Development Association - Results Measurement System, INDO-DAPOER, International Debt Statistics, Jobs for Knowledge Platform, LAC Equity Lab, Millennium Development Goals, Povstats, Quarterly Public Sector Debt, Quarterly External Debt Statistics/GDDS (New), Quarterly External Debt Statistics/SDDS (New), Readiness for Investment in Sustainable Energy (RISE), Sustainable Energy for All, Subnational Malnutrition Database, Subnational Poverty, Subnational Population, Wealth accounting, World Development Indicators and the Worldwide Governance Indicators.



is a simple task. The World Bank has a blog where they explain in detail how to use the APIs. See the web page *Accessing the World Bank Data APIs in Python, R, Ruby and Stata*, the web page *Welcome to wldata's documentation* and the web page *Package 'WDI'* for more details on how to do this. To put things in context, **Python** and **Ruby** are general-purpose programming languages, and **Stata** and **R** are programming environments optimized for statistics. They're all widely used in the business and academic worlds. The World Bank generates modules to those languages which help users to connect to the World Bank Development Indicators API and access the latest data.

For example, in **python**, the **wldata** module by Oliver Sherouse offers easy access to all the data in the World Bank's APIs. It also plays nicely with Wes McKinney's **pandas** analysis library<sup>4</sup>. **Wldata** is a simple python interface to find and request information from the World Bank's various databases, either as a dictionary containing full meta data or as a **pandas** Data Frame. Currently, **wldata** wraps most of the World Bank API, and also adds some convenience functions for searching and retrieving information.

In **R**, the **WDI** module by Vincent Arel-Bundock offers convenient access to the data in the World Bank's API. For fast searching, the **WDI** package ships with a local list of available data series. This local list can be updated to the latest version using the **WDIcache** function [web page *Package 'WDI'*]. Similar tools are available to languages such as Ruby and Stata.

Given that, the public data used in this project can be divided in two sets; first, the World Bank's World Development Indicators, which can be obtained by using the API and the packages mentioned; and, second, all the historical and major awards given by the World Bank to all developing countries from 1990 to 2014.

---

<sup>4</sup>See appendix B for **pandas** details.

### 3.2.1 World Development Indicators

The World Bank has a major data base called: World Development Indicators (WDI). WDI is the primary World Bank collection of development indicators, compiled from officially recognized international sources. This database represents the most current and accurate global development data available, and includes national, regional and global estimates. For the purpose of this work, the indicators selected such that prediction of corruption, collusion and fraud would not be attainable to specific country variables such as country or country region. This is was because the World Bank can not start investigation cases in specific countries just because different countries tend to have different levels of corruption. Unfortunately, those types of variables can have, in some times, much better prediction rates but the policies at the World Bank prevents a model of having them for discretion and discrimination arguments.

Also, the World Bank has around 16,000 different indicators the project could use, but being this big, and according to purpose of this project, only a few indicators were selected. The next list shows the ones that were considered. This list includes indicators related to the private sector and trade among countries which is the theme the project is targeting.

**IC.BUS.DISC.XQ** Private Sector & Trade: Business environment. Business extent of disclosure index (0=less disclosure to 10=more disclosure) Disclosure index measures the extent to which investors are protected through disclosure of ownership and financial information. The index ranges from 0 to 10, with higher values indicating more disclosure.

**IC.FRM.CMPU.ZS** Private Sector & Trade: Business environment. Firms competing against unregistered firms (% of firms).

**IC.FRM.CORR.ZS** Private Sector & Trade: Business environment Informal payments to public officials (% of firms).

**IC.FRM.INFM.ZS** Private Sector & Trade: Business environment Firms that do not report all sales for tax purposes (% of firms).

**IC.LGL.CRED.XQ** Private Sector & Trade: Business environment Strength of legal rights index (0=weak to 10=strong).

**IC.LGL.DURS** Private Sector & Trade: Business environment Time required to enforce a contract (days).

**IC.TAX.GIFT.ZS** Private Sector & Trade: Business environment Firms expected to give gifts in meetings with tax officials (% of firms).

**IQ.CPA.PROP.XQ** Public Sector: Policy & institutions CPIA property rights and rule-based governance rating (1=low to 6=high).

**IQ.CPA.TRAN.XQ** Public Sector: Policy & institutions CPIA transparency, accountability, and corruption in the public sector rating (1=low to 6=high).

**NY.GDP.PCAP.CD** Economic Policy & Debt: National accounts: USD at current prices: Aggregate indicators GDP per capita (current US\$).

**SE.PRM.PRSL.ZS** Education: Efficiency Persistence to last grade of primary, total (% of cohort).

**SI.POV.GINI** Poverty: Income distribution GINI index.

**SL.UEM.TOTL.NE.ZS** Social Protection & Labor: Unemployment Unemployment, total (% of total labor force) (national estimate).

The code to download WDI can be seen at the Appendix D in code D.1. This code can easily replicate results. For the purpose of this work, the code in D.1 downloads data for the countries within the countries from the investigations data, those countries define the countries list for this work.

As it can be seen from the figures 3.1 to 3.4, There are many differences among regions in the world. For example, on figure 3.1 that countries such as Vietnam in

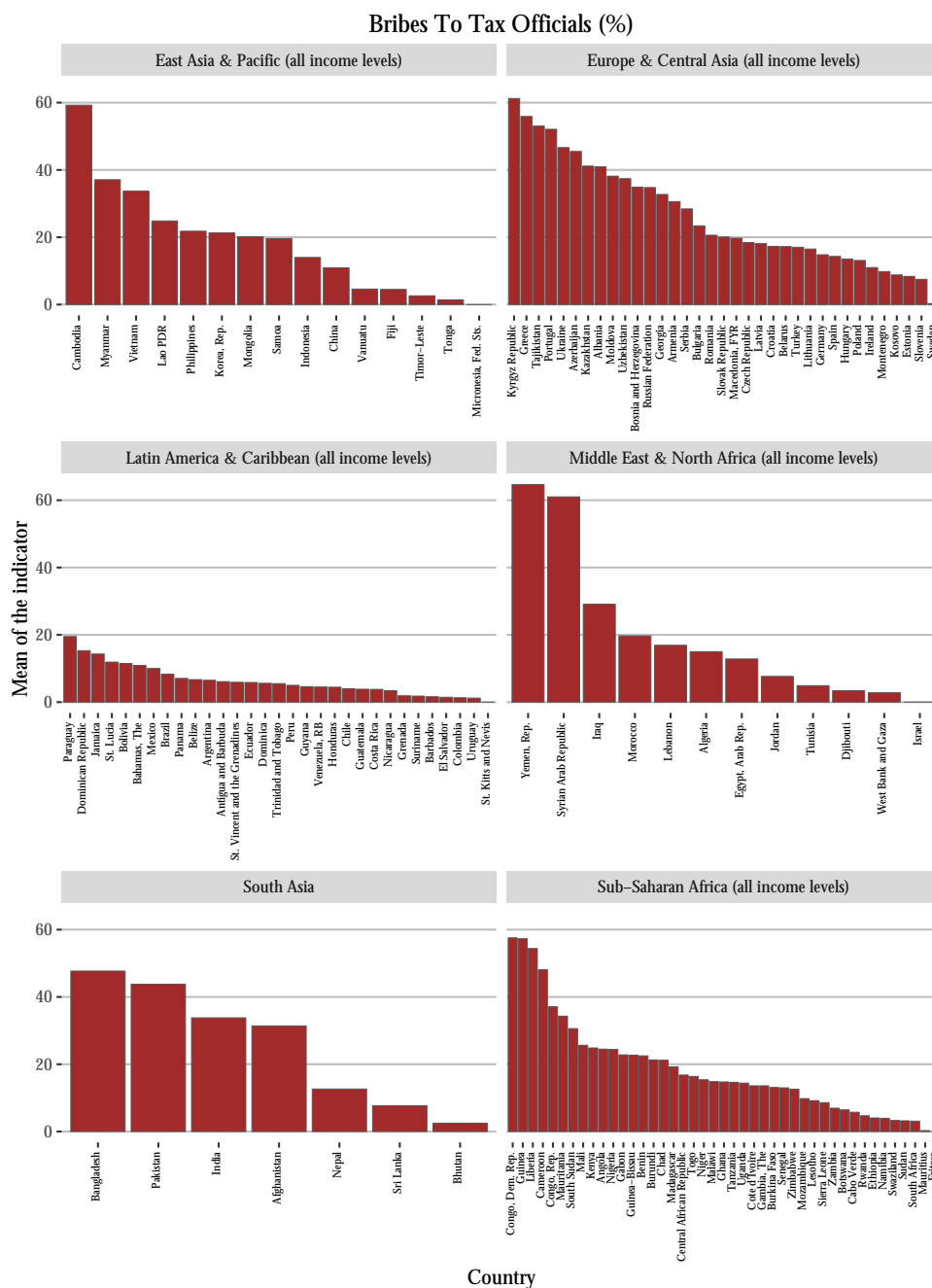
East Asia & Pacific, Grece, Portugal in Europe & Central Asia, Iraq in the Middle East & North Africa, Pakistan in South Asia and Democratic Republic of Congo, show high levels of percentage of Bribes to tax officials with levels higher than 40%. On the other hand, countries such as Micronecia in East Asia & Pacific, Slovenia, Sweden and Stonia in Europe & Central Asia, Chile, Mexico, Colombia and Uruguay in Latin America & Caribian, Israel and Tunisia in the Middle East & North Africa, Nepal and Sri Lanka in South Asia and South Africa and Rwuanda in Sub-Saharan Africa show low levels of percentage of Bribes to tax officials with numbers lower than 10%. Figure 3.2 shows the percentage of firms competing against informal firms. This indicator shows significant differences in each region. In the East Asia, Toga has the highest percentage near to 80% against Philippines and Lao PDR with just about 25%. In Europe and Central Asia, Kosovo has around 70% against Slovenia and Uzbekistan with 25%. In Latin America and the Caribbean, Bolivia, Uruguay and Mexico reach as far as 75% against Panama with levels near to 25%. In the Middle East and North Africa, Argelia has about 70% and Jordan and Israel just about 20%. In South Asia, Nepal and India have the highest levels with 50%, where Bhutan and Pakistan have around 25%. The highest levels are in Sub Saharan Africa where countries such as Sudan, Chad, Cameroon and Niger reach up to 80% against Eritrea with levels higher than 25%. Now, figure 3.3 shows the percentage of Payments to public officials where clearly, one more time, countries in Sub Saharan Africa have the highest levels like Guinea, Democratic republic of Congo, Republic of Congo and Cameroon reach up to 80%. Countries like Syrian Arab Republic and Argelia in the Middle East, Kyrgyz Republic and Ukraine in Europe and Central Asia, Cambodia and Vietnam in East Asia, Paraguay in Latin America and Bangladesh in South Asia present also high levels near to 65% in contrast with countries such as Sudan and Cabo Verde in Sub Saharan Africa, Israel and Tunisia in the Middle East and North Africa, Spain, Ireland and Montegro in Europe and Central Asia, Fiji, China and Tonga in East Asia and the Pacific, Mexico, Uruguay and Colombia in Latin America and Caribbean, India and Nepal in South Asia, present very low levels near to 20%.

Finally, figure 3.4 shows the mean time to enforce contracts in each country in hours. Differences are very radical. In East Asia, Timor-Leste has the highest number near to 1500 hours (63 days) against Singapore and New Zealand with just about 200 hours (9 days). In Europe and Central Asia, Slovenia and Italy it takes about 1300 (55 days) against Russian Federation, Finland and Luxemburg where it takes about 180 (8 days). In Latin America and the Caribbean, Suriname, Guatemala and Colombia have the longer time with about 1400 hours (58 days) against Mexico and Antigua & Barbuda near to 350 (14 day). In the Middle East and North Africa, in Djibouti and Egypt Arab Republic it takes near to 1200 hours (50 days) where in Malta, Morocco and Iraq just takes about 500 hours (20 days). As a reference, in USA and Canada it takes about 500 hours (20 days). In South Asia in countries such as Afghanistan and Bangladesh it takes about 1500 hours (63 days) and in Bhutan just about 250 (10 days). In the Sub Saharan Africa, the record of highest time goes to Guinea - Bissau with times up to 1700 hours (70 days) and the lower to South Sudan with just about 250 hours (10 days).

Appendix C show the figures of all the other indicators used in this project obtained with the package WDI [web page *Package ‘WDI’*] and the code from appendix D.

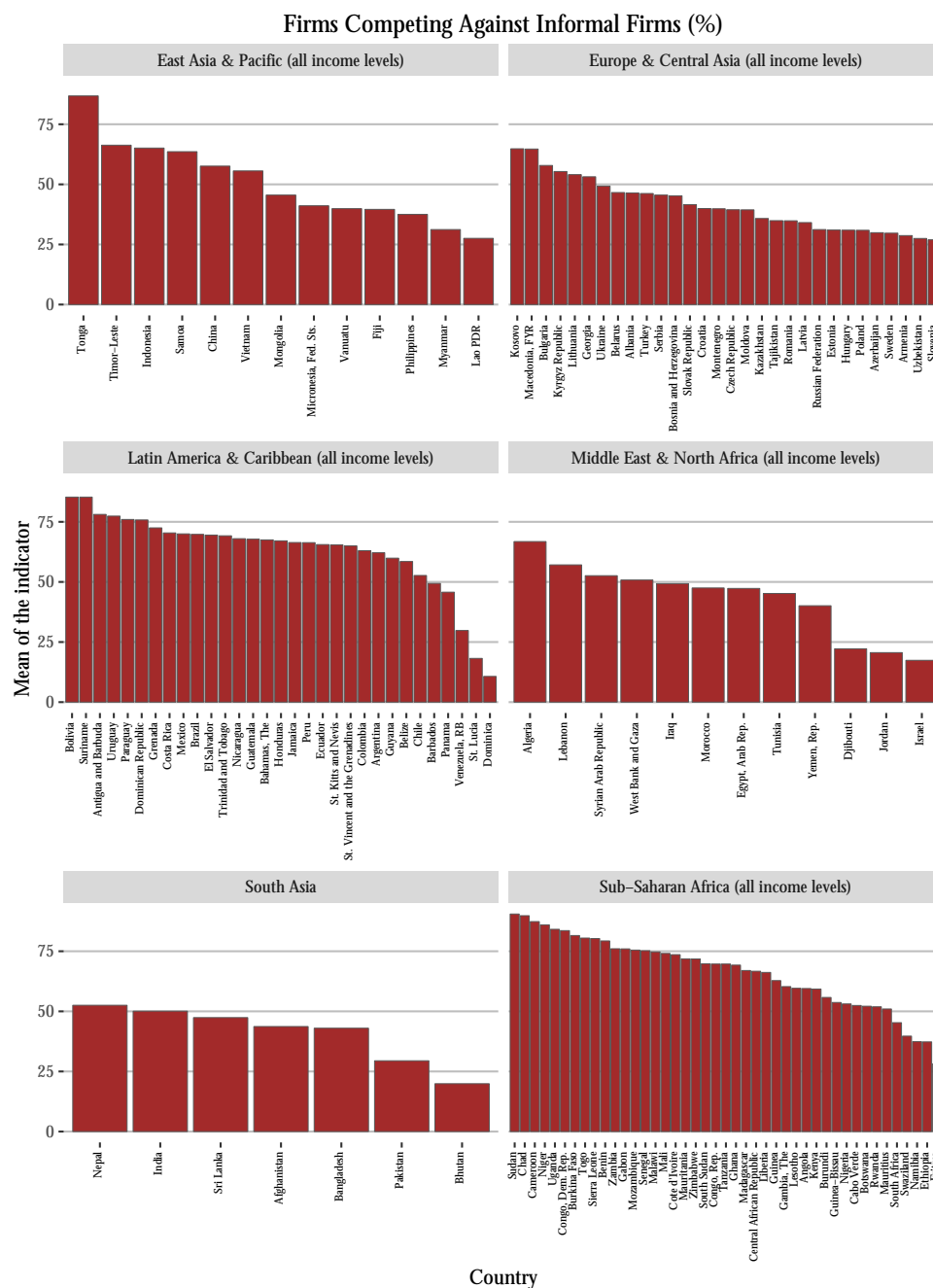
These indicators were chosen because each one of them suggest that variances within them show an impact on how many contracts present substantiated cases of corruption, collusion, coercion and fraud among the contracts that the World Bank have given to development projects in those countries. As chapter 2 says and figures 2.2a to 2.2g suggest, differences in the indicators among countries and regions might be clues to identify malicious contracts.

Figure 3.1. WDI: % of Bribes to tax officials per Country/Region



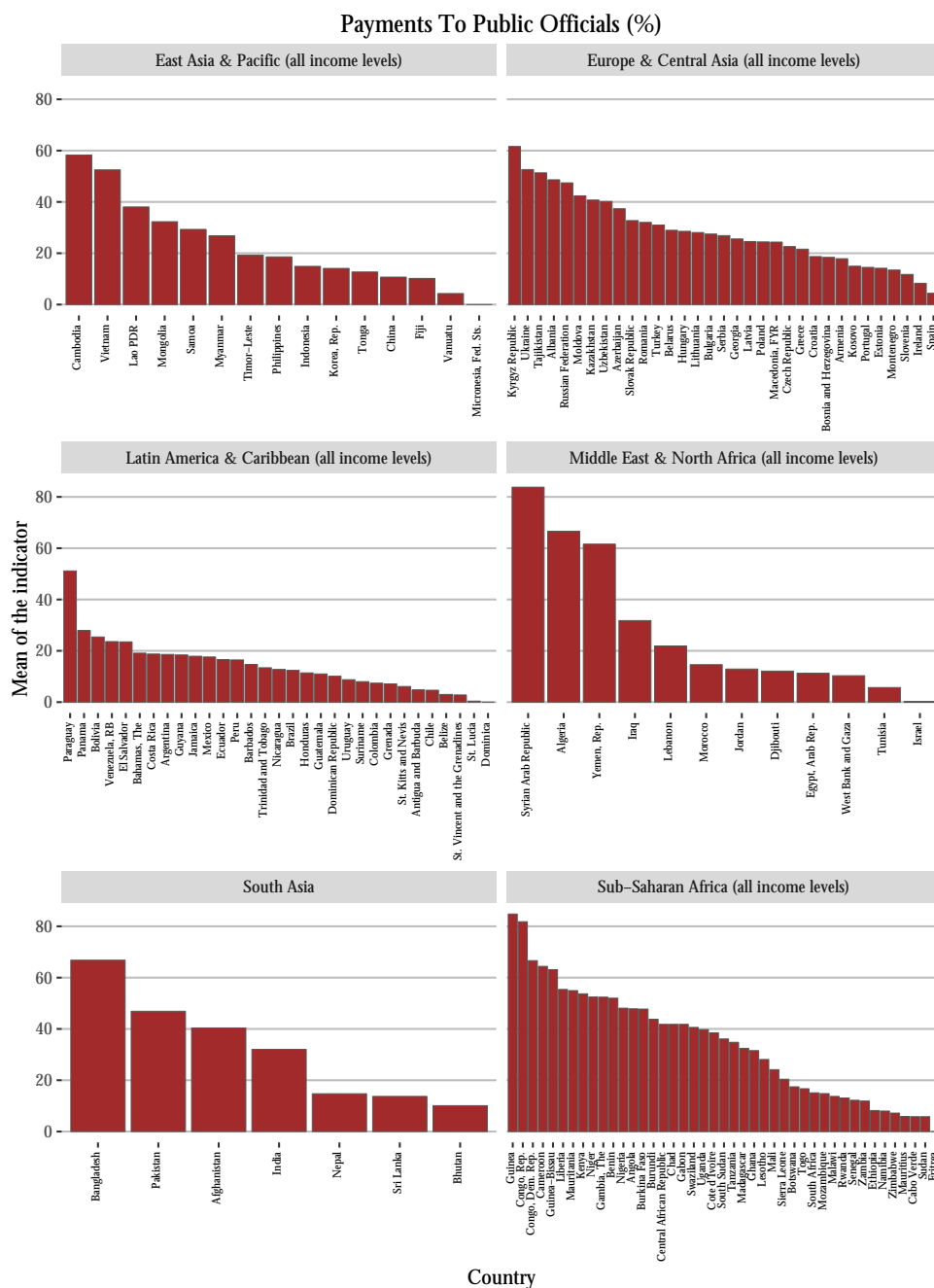
**Source:** Own creation and data obtained with package WDI, see the web page *Package 'WDI'*.

Figure 3.2. WDI: % of Firms competing against informal firms per country/Region



**Source:** Own creation and data obtained with package WDI, see the web page *Package 'WDI'*.

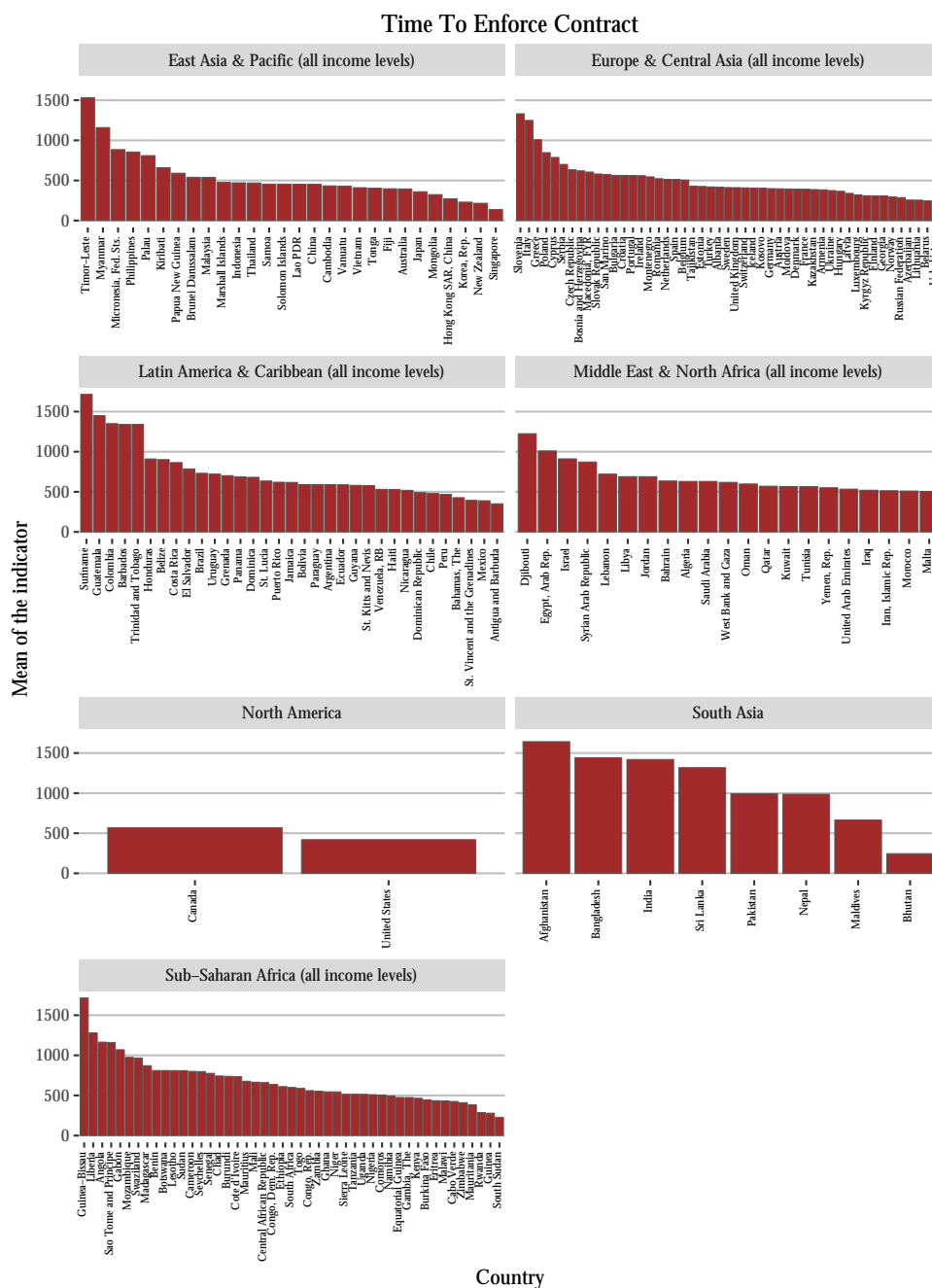
Figure 3.3. WDI: % of Payments to public officials per Country/Region



**Source:** Own creation and data obtained with package WDI, see the web page *Package 'WDI'*.



Figure 3.4. WDI: Time to Enforce Contracts per Country/Region



**Source:** Own creation and data obtained with package WDI, see the web page *Package 'WDI'*.

### 3.2.2 Major & Historic Awards

As it was explained previously, the other main public dataset was the Major and Historic contracts awards from the World Bank. The data include the variables shown in table 3.1. The variables shown refers to the names after an extensive cleaning process that involved merging data from all the Development Banks' sites and from the Historical Awards data obtained from the World Bank. Such variables include Procurement category that refers to whether if the procurement was of Consultant Services, Non-Consultant Services, Goods or Civil Works. The type of agreement refers to International Bank for Reconstruction and Development (IBRD), Recipient-Executed Trust Funds (RETF), International Development Association (IDA), Institutional Development Fund (IDF), Global Environment Facility (GEF), State and Peace-Building Fund (SPF), California Air Resource Board (CARB) or Debt Reduction Facility (DRF) The Bid type can be Quality And Cost-Based Selection, Single Source Selection, Selection Based On Consultant's Qualification, Individual, National Competitive Bidding, International Competitive Bidding, CQS, SHOP, Direct Contracting, Quality Based Selection, Least Cost Selection, Selection Under a Fixed Budget, National Shopping, International Shopping or Limited International Bidding

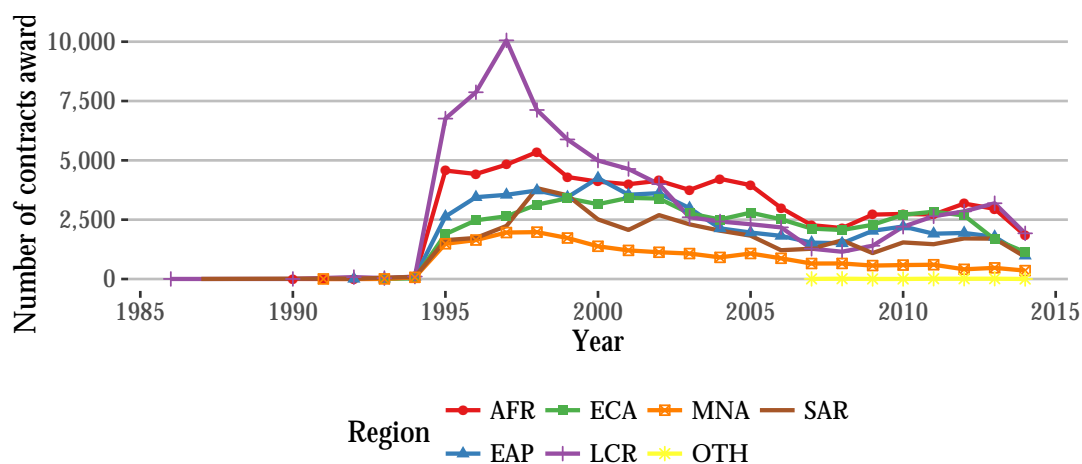
The historical and mayor awards data has information of contracts awarded between 1986 to 2014. As it can be seen on figure 3.5 the historical awards correspond to years from 1986 to 1994 and the major awards are the ones from 1995 to June, 2014. Figure 3.5 shows that the number of contracts have been reduced in the recent years in every region of the world and there was an abnormal amount of contracts awarded in the Latin America and Caribbean region between 1996 and 1999. Figure 3.6 shows the total awarded amount of the Historical and Major contracts by each specific region. This figure shows that although the number of contracts was very high in the Latin America and Caribbean region in 1996 to 1999, the total awarded money to that region in US dollars was not as high as the number of contracts.

Table 3.1. Major and Historic Awards Dictionary

Variable	Values
agreement_type	Agreement type
award_amount_usd	Numeric
award_currency	USD
award_date	Date
bider_country_2nd	Country
bider_country_3rd	Country
bider_country_4th	Country
bid_type	Type of Bid
buyer	Name of the contractor
buyer_country	Country
buyer_country_id	Country ID
buyer_reference	Identifier for the project
competitive	Logical
contract_description	Description of the project
domestic_pref_affect	Logical
domestic_pref_allowed	Logical
fiscal_year	Date
ln_cr_id	Loan ID
major_sector	Sector
number_of_bids	Numeric
number_of_firms	Numeric
number_of_suppliers_awarded	Numeric
price_escalating_flag	Logical
procurement_category	Procurement Category
product_line_original	Logical
project_id	Project ID
project_name	Project Name
region	ECA, AFR, SAR, LCR, MNA, EAP, OTH
sent_date	Sent date
signing_date	Sign date
supplier	Supplier name
supplier_country	Supplier Country
supplier_country_id	Supplier Country ID
unique_id	Unique ID
wb_contract_number	Contract Number
historical	Logical
country	Country
year	Year

**Source:** Own creation.

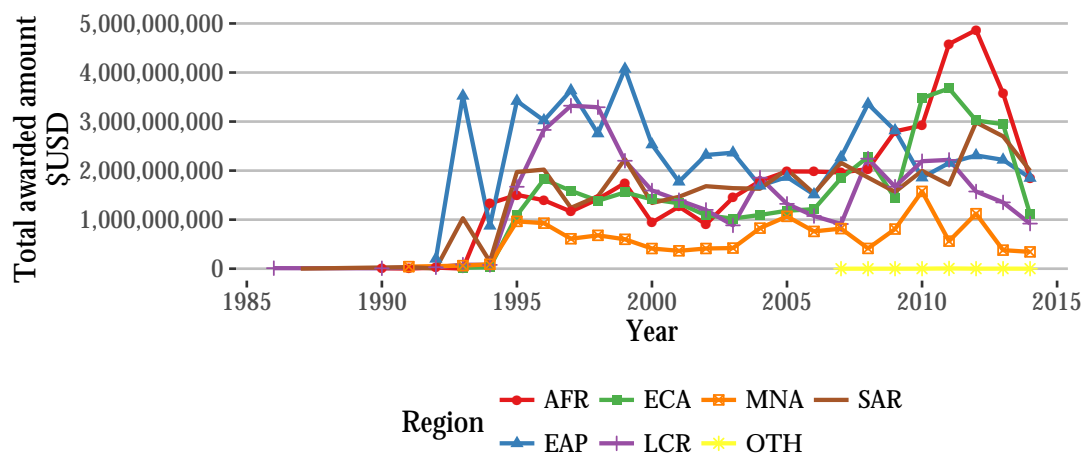
Figure 3.5. Number of Historic &amp; Major awards per Region



**Source:** Own creation with data from the World Bank.

**Note:** OTH refers to projects that cannot be assigned to a specific country. Data to June, 2014.

Figure 3.6. Total awarded amount of Historic &amp; Major contracts per Region



**Source:** Own creation with data from the World Bank.

**Note:** OTH refers to projects that cannot be assigned to a specific country. Data to June, 2014.

Figure 3.6 also shows that just for 2013, the total awarded amount adds up to around 13,200 billion US dollars<sup>5</sup> and that the awarded amount has increased more in Africa and in Europe from 2009 to 2013. It's important to notice that the data goes up to June, 2014 so the fall in the last point in figures 3.5 and 3.6 of each series for each region is not necessarily going down, it is probably better to ignore that data.

Now, figure 3.7 in page 43 shows the top six countries in terms of awarded contracts. It is interesting how tendencies show the World Bank's priorities among time. For example, it is clear that the top countries from 2008 to 2015 has been Vietnam in first place, Afghanistan in second, and China and some Latin American countries in the next places. From 1998 to 2006, the tendencies were different, China and India take the first and second places with around 1,500 contracts per year for that period. In the next places, countries such as Brazil, Bosnia Herzegovina, Argentina and the Russian Federation. Finally, in the term from 1990 to 1997, Mexico, Argentina and India take the first places followed by countries such as China, Bolivia, Ecuador, and other Latin American countries.

Next, figure 3.8 in page 44 shows the top six countries in terms of total awarded amount for each year in US dollars. From the figure, it is easy to see that China has been the country that has received the higher amount of money in contracts followed by India in second place and Brazil in third. This tend seems to be a constant in the top three countries since 1988 up to 2014. Other countries in the next three places include the Russian Federation, Vietnam, Africa and Africa.

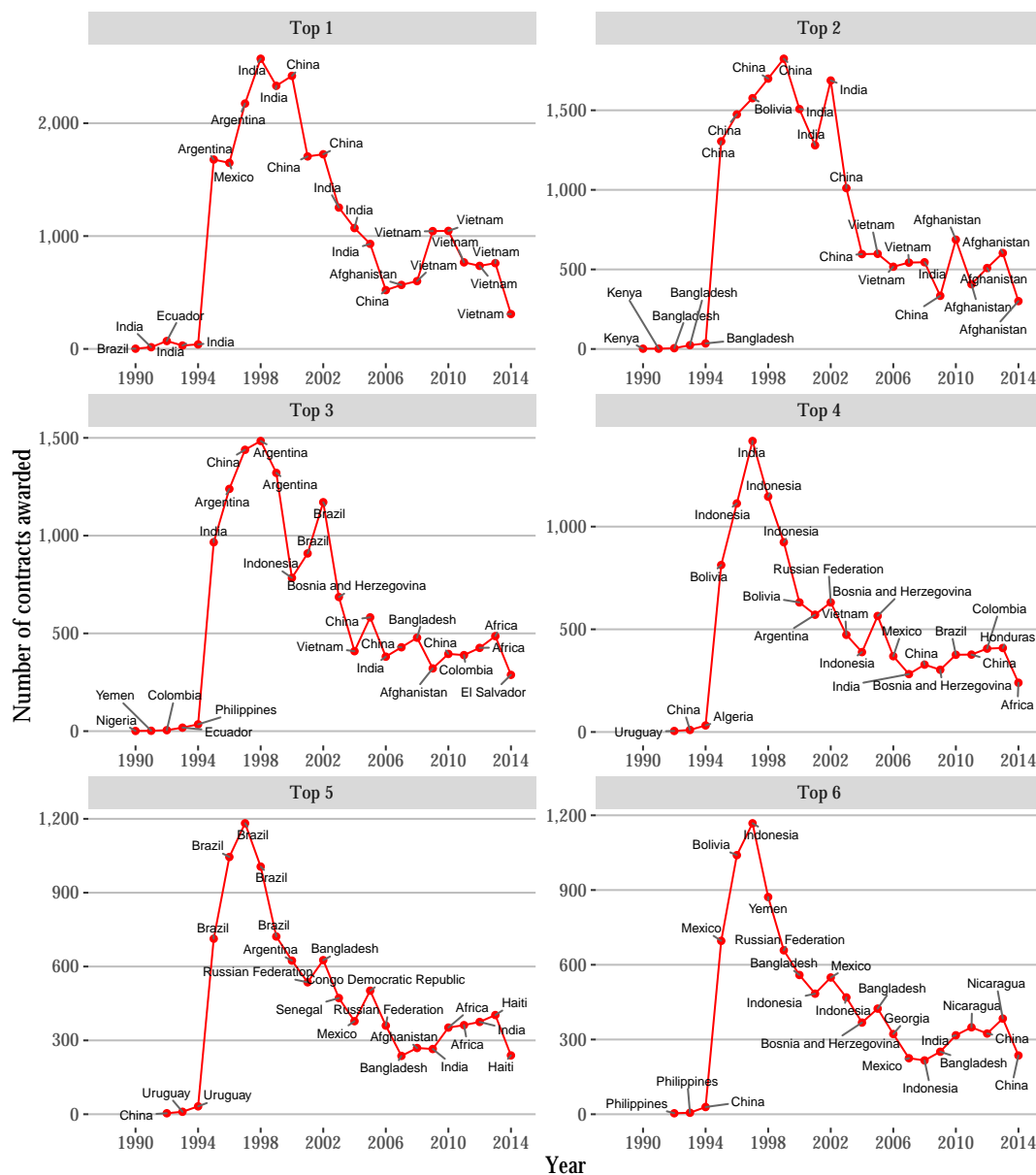
The visualizations of the data from figures such as 3.5 to 3.8 could help the World Bank to prioritize how to administer the number of contracts to award and the total awarded amount among different regions in the world and among different time period such that there's a balanced assignation of resources because figures 3.5 to 3.8 suggest that the World Bank is not taking that into their assignation policy. This fact might or might not be related to how corruption, collusion and fraud but

---

<sup>5</sup>1 billion = 1,000,000,000.

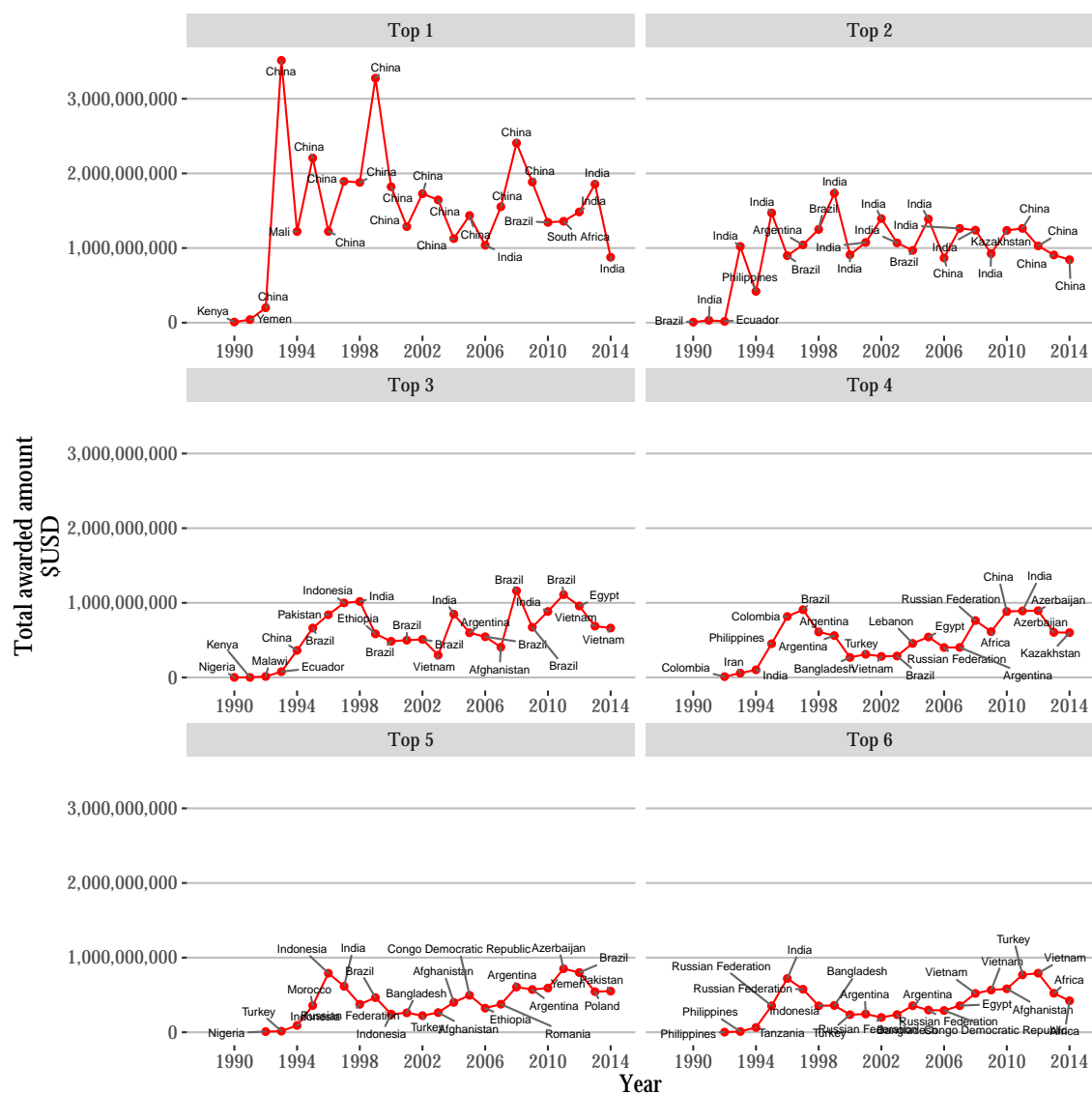
it certainly helps them as a political strategy.

Figure 3.7. Top six countries: number of awarded contracts



Source: Own creation with data from the World Bank.

Figure 3.8. Top six countries: total awarded amount (\$USD)



Source: Own creation with data from the World Bank.

### 3.3 Entity names disambiguation

As it was explained on chapter 2, historically, the World Bank has made investigations based, mainly, on information and data provided third parties claiming that something is suspicious within a procurement process. Fortunately, the World Bank has been keeping record of the investigation process by generating the investigations data. The problem is that the information on the investigations data can not be easily linked to the Major and Historical awards since the suppliers, the bidders and the contractors' names were introduced in many different formats by many different organizations among the World Bank as described in chapter 1. This is a current limitation for the World Bank and a big problem regarding this project. This section explains one way to approach this problem.

The idea is very simple. Technology in modern search engines such as Google provides a reliable way to search for each one of the names registered in the data from the Investigations and the Major & Historical awards data. So the plan was to *google* each one of the unique names among the different data sources and compare the results so that entities that have similar results correspond to the same entity. This is a major data problem because it escalates very fast. After *googling* each one of the entities, results must be compared with all the remaining results in order to determine whether they are considered similar or not. This means that if  $n$  searches are made, then  $n^2$  comparatives has to be computed.

The first step then became very clear, removing words such that the problem could be simplified. This was referred as stop words removal. The idea behind removing the stop words is to simplify the work so that companies or entity names that are supposed to refer to the same organization or person but have different names among the data sources such as Price Waterhouse Coopers incorporated and Price Waterhouse Coopers INC actually do. By doing so, the entity names was reduced in about 15% passing from 234,067 entity names originally to 198,957 unique values after removing punctuation and stop words, which is a very considerable amount



taking into account that the idea is to search every one of them in Google.

The first problem then becomes that unique names from contractors, buyers and bidders ascend up to 198,957 unique names in the different data sources after cleaning for stop words and removing punctuation<sup>6</sup>. This has two main limitations, the first that, Google, does not allows users to perform computed searches that ascend to a number as big as what we are dealing here. The second, even if we were able to perform those searches, the number of binary comparatives to be made after the searches will be a number near to 39,204,000,000 which, if each binary computation lasted a 10% of a second, it would take about 124 years to complete all of them sequentially, and, of course, that is not useful at all.

Of course, the plan was not to perform sequential operations to make all the comparatives because, even if we had a 124 machines cluster to compute that in parallel, which is expensive enough, that process would still take about a year to end.

So the algorithm to approach this problem was the following: first, divide all the canonical names into several chunks and start an AWS machine, each with different I.P. address so that we don't face the restriction of making too much automated searches at Google; second, perform the searches for each one of the entity names and store the first 10 results given by google; third, after having the results within each chunk, compute the binary comparisons in that chunk to see how many searches in common entities have and if any two entities have the desired number of search results in common, group them as one company and then tag just one of them so that in the next step, fourth, the final step, just make the binary comparatives with the entities that had no matches and the ones tagged so that the process reduces the number of comparisons to be made. The code to perform this process is displayed in the appendix D code D.2 and D.3

---

<sup>6</sup>The stop words were selected based on insight form the Integrity Vice Presidency. The stop words used were: 'jsc', 'corp', 'm/s', 'inc', 'incorporated', 'doo', 'd.o.o.', 'la', 'el', 'srl', 's.r.l.', 'limited', 'llc', 'l.l.c.', 'co', 'corporation', 'company', 'ltd', 'ltda', 'lda', 'de', 'i/e', 'ooo', 'limited', 'societe', 'société', 'et', 'le', 'and', 'the', 'of', 'de', 'des', 'du', 'sa', 's.a.', 'pty', 'ste', 'sarl', 's.a.r.l.', 'gmhb', 'g.m.h.b.', 'sprl', 's.p.r.l.'.

Results were great because this algorithm completed in less than four days using about 450 instances. That means that a process that would normally take at least 100 days (by using 450 instances) was reduced in about 95% the time making it very efficient. This was a great result because it created very useful results in a decent time for the project. Next subsection covers how to decide the parameters used for this process.

### 3.3.1 Entity names disambiguation parameters & results

The results of the algorithm described in the last section and displayed in the appendix D code D.2 and D.3 depend on what is considered to be a common entity name in terms of results in the Google searches. In other words, it depends on how many results of the Google searches in common are needed in order to consider two entities to be the same one.

Approaching this represents a trade-off between lowering the error in the number of unique entity names (false negatives) against being sure that two entities are actually the same in reality (false positives). For this project it was more relevant to reduce the false positives to be sure that whenever two entities have the same canonical name it's actually the case.

In order to decide how many searches in common were needed, we performed several experiments and talked with the partners at the Integrity Vice Presidency from the World Bank. They agreed that we needed a conservative approach to this problem. We tried using nine, seven, five and three result searches in common within a random sample so that we could compare results in a viable time frame. Results using nine and seven searches in common seemed to be too conservative. While, results using three results seemed desirable because it reduced the entity names in about 14%, it seemed too relaxed for our partner at the World Bank so we decided to stay with five searches in common. The total number of unique names was reduced in about 10% which left us with great data for the next steps of the model.

The output of this algorithm was the first result of the project and this alone

represented a first way to better understand procurement data among the different databases of the World Bank. Partners at the Integrity Vice Presidency were surprised that the project came up with a useful method to reconcile data that the investigators at the World Bank can actually use. The project itself could be just about a writing the best algorithm to reconcile entities among different databases that's why the investigators from the World Bank were thanked enough just for this results. It is very important to say that the output of this process is and will be used just as an additional tool that helps them conduct their investigation process and in no way they are considering that the results of this process are absolute in any form.

Just to state a very clear example of how this algorithm worked, table 3.2 shows all the suppliers that were matched to the entity Price Waterhouse Coopers.

Table 3.2. Entity names disambiguation example

Supplier	Canonical Name
PRICE WATERHOUSE AUDITORES	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE/NORCONSULT	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS (PWC)	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS LIMITED	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS LTD.	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSECOOPERS PVT. LTD.	PRICE WATER HOUSE COOPERS
MCO COOPERATING - PRICE WATER HOUSE	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS CONSULTANTS SINGAPORE PTE LTD	PRICE WATER HOUSE COOPERS
SIR WILLIAM HALCROW/PRICE WATERHOUSE PAR	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSECOOPERS ASESORES GERENCIALES LTDA.	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS DEL ECUADOR	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE S.C./TP/MA	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE - AFRICA	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS ASSOCIATES AFRICA LTD.	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE EASTERN EUROP	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE/PR. WAT. & CO	PRICE WATER HOUSE COOPERS
A/O PRICE WATERHOUSE/PRICE WAT	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE LLP, VA. USA	PRICE WATER HOUSE COOPERS

Continues on next page

Table 3.2. Entity names disambiguation example

Supplier	Canonical Name
PRICE WATERHOUSE - GMI AUDIT	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE/COOPERS	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE MEYER NEL	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE PAN AFRICAN CONSULTANTS	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS AND HOWARD HUMPHREYS	PRICE WATER HOUSE COOPERS
PISTRELLI, PRICE WHSE COOP&LYB	PRICE WATER HOUSE COOPERS
CENIT PRICE WATERHOUSE	PRICE WATER HOUSE COOPERS
GMS SA - PRICE WATER	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE & CO.	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE Y CO.,S.C.	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS CONSULTORES S.A.	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS Y COMPANIA, S.C.	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE INTERAMERICAN	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS LLC	PRICE WATER HOUSE COOPERS
CONSORTIUM ROTHSCHILD/LANDWELL & ASSOCIATES/PRICE WATERHOUSECO	PRICE WATER HOUSE COOPERS
PW-PRICE WATERHOUSE	PRICE WATER HOUSE COOPERS
PRICE WATER HOUSE COOPERS	PRICE WATER HOUSE COOPERS
PRICE WATERHOUS COOPERS	PRICE WATER HOUSE COOPERS
PLANT LOCATION INT'L/PRICE WA.	PRICE WATER HOUSE COOPERS
BNP(PARIS)PRICE WATERHOUSE	PRICE WATER HOUSE COOPERS
PRICE WATER HOUSE COOPERS LLP	PRICE WATER HOUSE COOPERS
ROCHE/PRICE WATER HOUSE	PRICE WATER HOUSE COOPERS
ROCHE INTER./PRICE WATERHOUSE	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPER LTD.	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS LAO LTD	PRICE WATER HOUSE COOPERS
SAMIL PRICE WATERHOUSE COOPERS	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSECOOPERS	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPER VIETNAM	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE LLC	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE,SOFIA,BULGARI	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS KFT	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE LLP	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS LLP	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS, UZBEK BRANCH	PRICE WATER HOUSE COOPERS
PRICE WATER HOUSE	PRICE WATER HOUSE COOPERS

Continues on next page

Table 3.2. Entity names disambiguation example

Supplier	Canonical Name
PRICE WATERHOUSE S.A.,	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE/SABAS APPROTECH PROJECT	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE/GONZ VILCHIS	PRICE WATER HOUSE COOPERS
GONZALEZ VILCHIS/PRICE WATERHOUSE	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS N.V.	PRICE WATER HOUSE COOPERS
PRICE WATERHOUSE COOPERS ANTIGUA	PRICE WATER HOUSE COOPERS

End of table

**Source:** Own creation with data from the World Bank.

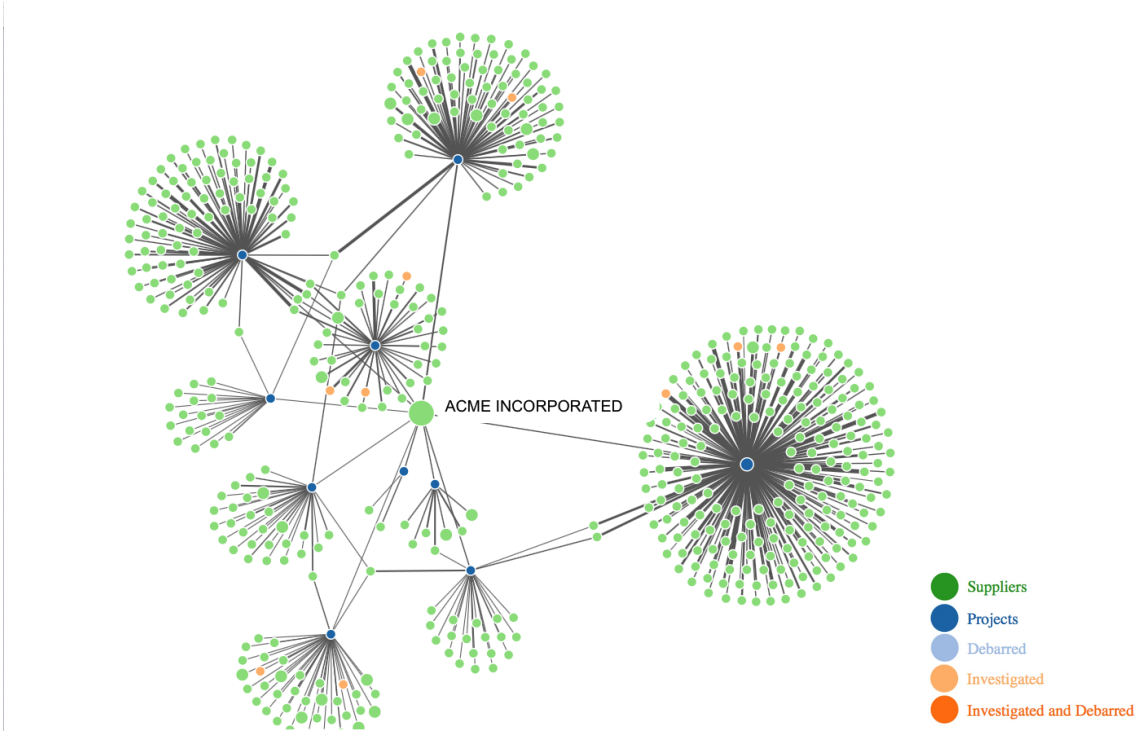
### 3.4 Co-Award Network and feature creation

After waiting for this algorithm to end, we finally were ready to build tools for a proactive investigation within the World Bank. To evaluate contract risk, we generated features and models tracking companies' historical involvement on World Bank projects within specific countries and sectors. As well we created co-award network features for each company. Those features would be used later in the model and estimations. Features facilitate the identification of patterns that could be related to potential corruption, collusion, coercion and fraud.

For example, down here at figure 3.9 you can see the co-award network for ACME INCORPORATED, a fictitious entity created to illustrate how does the co-award network works. In dark blue there's every project that the ACME INCORPORATED has been part of and in green there's every company that worked in that specific project. This is the private version of the network delivered to the Integrity Vice Presidency in the web application (for more details see 4). This co-award network is a level one network that includes they have different colors whether a company was investigated and found to be guilty, white blue represents a company or an entity that have been previously debarred from receiving future projects from the World Bank. Light orange represents a company or an entity that have been investigated in the past and, finally, darker orange represents a company or an entity that have been investigated and debarred. As explained in chapter 2, there can be many reasons for the World Bank to grant debarment for a company or an entity. The important thing to consider here is that the fact that an entity worked on a project where there's one or more entities that have been investigated, debarred or both, might be an indication of collusion, corruption or even fraud so the idea is to incorporate those features to the data so that the model

is able to identify those potential indicators if they actually exist.

Figure 3.9. Co-award Network example



**Source:** Own creation with data from the World Bank.

Table 3.3 show all the features created from the co-award network.

Table 3.3. Co-award network features

Features
Supplier_Degree_Centrality_Contemporary_Global
Project_Degree_Centrality_Contemporary_Global
Supplier_Degree_Centrality_Cumulative_Global
Project_Degree_Centrality_Cumulative_Global
Supplier_In_Giant_Component_Contemporary_Global
Project_In_Giant_Component_Contemporary_Global
Supplier_In_Giant_Component_Cumulative_Global
Project_In_Giant_Component_Cumulative_Global
Supplier_Neighbor_Intensity_Contemporary_Global
Project_Neighbor_Intensity_Contemporary_Global

Continues on next page

Table 3.3. Co-award network features

Features
Supplier_Neighbor_Intensity_Cumulative_Global
Project_Neighbor_Intensity_Cumulative_Global
Supplier_Neighborhood_Size_k_Contemporary_Global
Project_Neighborhood_Size_k_Contemporary_Global
Supplier_Neighborhood_Size_k_Cumulative_Global
Project_Neighborhood_Size_k_Cumulative_Global
Supplier_Betweenness_Centrality_Contemporary_Global
Project_Betweenness_Centrality_Contemporary_Global
Supplier_Betweenness_Centrality_Cumulative_Global
Project_Betweenness_Centrality_Cumulative_Global
Supplier_Minimum_Distance_Investigated_Suppliers_Contemporary_Global
Project_Minimum_Distance_Investigated_Suppliers_Contemporary_Global
Supplier_Average_Distance_Investigated_Suppliers_Contemporary_Global
Project_Average_Distance_Investigated_Suppliers_Contemporary_Global
Supplier_Number_in_Neighborhood_Size_k_Investigated_Suppliers_Contemporary_Global
Project_Number_in_Neighborhood_Size_k_Investigated_Suppliers_Contemporary_Global
Supplier_Minimum_Distance_Investigated_Suppliers_Cumulative_Global
Project_Minimum_Distance_Investigated_Suppliers_Cumulative_Global
Supplier_Average_Distance_Investigated_Suppliers_Cumulative_Global
Project_Average_Distance_Investigated_Suppliers_Cumulative_Global
Supplier_Number_in_Neighborhood_Size_k_Investigated_Suppliers_Cumulative_Global
Project_Number_in_Neighborhood_Size_k_Investigated_Suppliers_Cumulative_Global
Supplier_Minimum_Distance_Investigated_Projects_Contemporary_Global
Project_Minimum_Distance_Investigated_Projects_Contemporary_Global
Supplier_Average_Distance_Investigated_Projects_Contemporary_Global
Project_Average_Distance_Investigated_Projects_Contemporary_Global
Supplier_Number_in_Neighborhood_Size_k_Investigated_Projects_Contemporary_Global
Project_Number_in_Neighborhood_Size_k_Investigated_Projects_Contemporary_Global
Supplier_Minimum_Distance_Investigated_Projects_Cumulative_Global
Project_Minimum_Distance_Investigated_Projects_Cumulative_Global
Supplier_Average_Distance_Investigated_Projects_Cumulative_Global
Project_Average_Distance_Investigated_Projects_Cumulative_Global
Supplier_Number_in_Neighborhood_Size_k_Investigated_Projects_Cumulative_Global
Project_Number_in_Neighborhood_Size_k_Investigated_Projects_Cumulative_Global

End of table

---

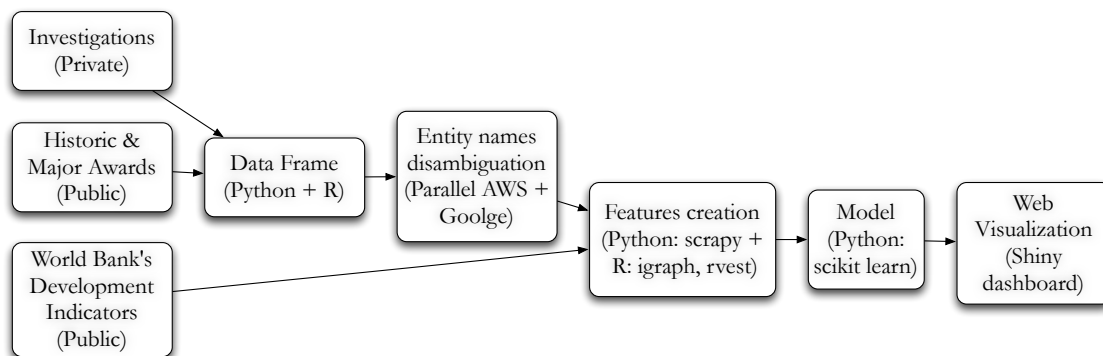
## Chapter 4.

# Detecting corruption, collusion & fraud: Data product

---

### 4.1 Data pipeline

Figure 4.1. Data pipeline



**Source:** Own creation.



## 4.2 Red-flags from data

Chapter 2 presented figure 2.2 suggesting ideas of how to look for common patterns to identify potential cases of corruption, collusion, fraud, coercion and other types of malicious behavior among the procurements that the World Bank gives to countries all over the world. The objective of this section is to try to replicate those figures by using real data in order to provide a valuable insight to the investigators working in the Integrity Vice Presidency to help them attack this problem.

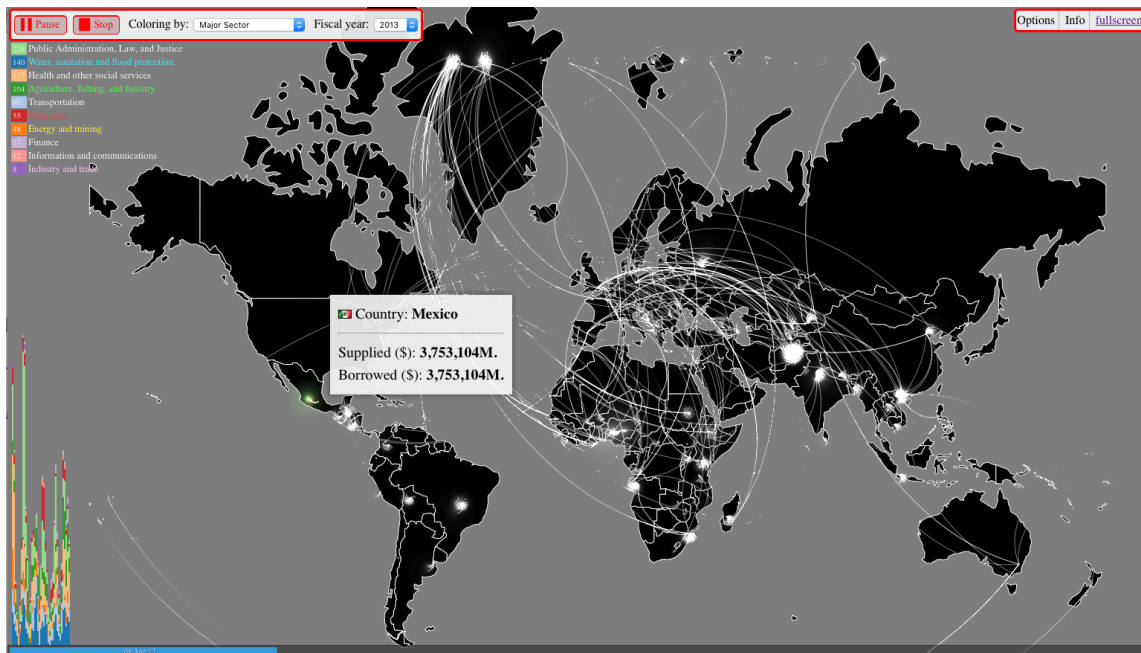
## 4.3 Model

## 4.4 Web visualization application: dashboard

### 4.4.1 Dashboard outline

### 4.4.2 Interactive map

Figure 4.2. Interactive map



**Source:** Own creation based on web page *World Bank Contract Awards*.

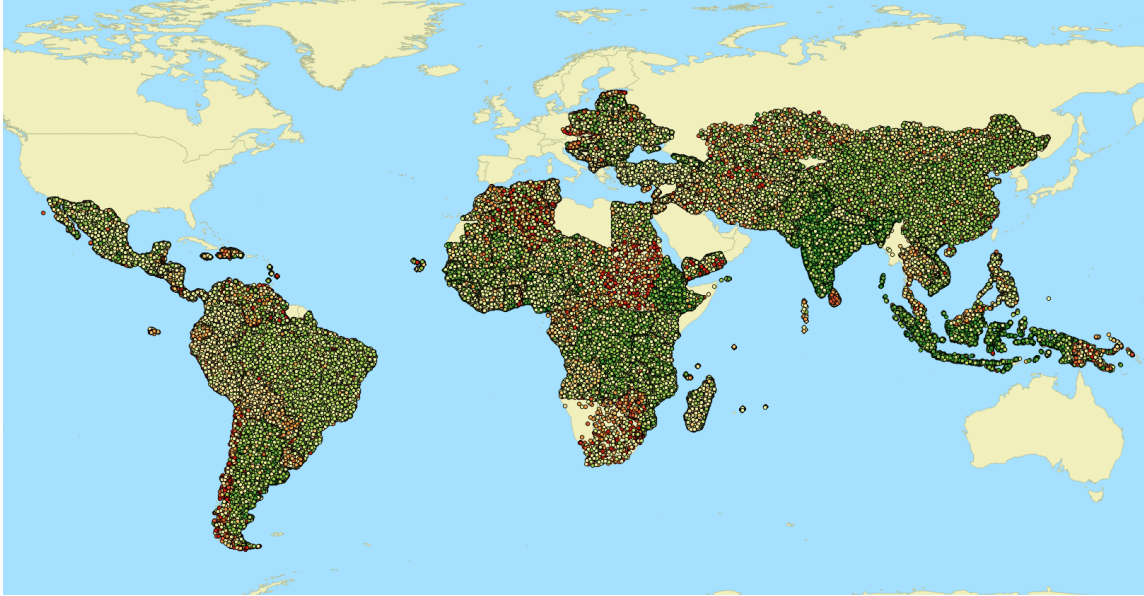
Go to [detecting-corruption.carlospetricioli.com/interactive\\_map](http://detecting-corruption.carlospetricioli.com/interactive_map) to see a live version.

Data from the World Bank [web page *Data Bank*].

#### 4.4.3 Companies & projects network

#### 4.4.4 Contract specific risk map

Figure 4.3. Risk map (Sample)



**Source:** Own creation based on web page *World Bank Contract Awards*.  
Data from the World Bank [web page *Data Bank*].

---

## Appendix A.

### Sample Procurement Plan

---

*(Text in italic font is meant for instruction to staff and should be deleted in the final version of the PP)*

*(This is only a sample with the minimum content that is required to be included in the PAD. The detailed procurement plan is still mandatory for disclosure on the Bank's website in accordance with the guidelines. The initial procurement plan will cover the first 18 months of the project and then updated annually or earlier as necessary).*

#### I. General

1. **Bank's approval Date of the procurement Plan** *[Original: December 2007]: Revision 15 of Updated Procurement Plan, June 2010]*
2. **Date of General Procurement Notice:** *Dec 24, 2006*
3. **Period covered by this procurement plan:** *The procurement period of project covered from year June 2010 to December 2012*

#### II. Goods and Works and non-consulting services.

1. **Prior Review Threshold:** Procurement Decisions subject to Prior Review by the Bank as stated in Appendix 1 to the Guidelines for Procurement: *[Thresholds for applicable procurement methods (not limited to the list below) will be determined by the Procurement Specialist /Procurement Accredited Staff based on the assessment of the implementing agency's capacity.]*

Table A.1. Prior Review Threshold

	Procurement Method	Prior Review Threshold US\$	Comments
1.	ICB and LIB (Goods)	<i>Above US\$ 500,000</i>	<i>All</i>
2.	NCB (Goods)	<i>Above US\$ 100,000</i>	<i>First contract</i>
3.	ICB (Works)	<i>Above US\$ 15 million</i>	<i>All</i>
4.	NCB (Works)	<i>Above US\$ 5 million</i>	<i>All</i>
5.	(Non Consultant Services) [Add other methods if necessary]	<i>Below US\$ 100,000</i>	<i>First contract</i>

**Source** web page *Procurement Plan Template*.

2. **Prequalification.** Bidders for *Not Applicable* shall be prequalified in accordance with the provisions of paragraphs 2.9 and 2.10 of the Guidelines.
3. **Proposed Procedures for CDD Components (as per paragraph. 3.17 of the Guidelines):** *[Refer to the relevant CDD project implementation document approved by the Bank or delete if not applicable]*
4. **Reference to (if any) Project Operational/Procurement Manual:** *Project Implementation Manual for World Bank Loan Project XYZ 04/01/2010 issued by < mention name of PIU>*
5. **Any Other Special Procurement Arrangements:** *[including advance procurement and retroactive financing, if applicable] 5 ICB works packages will be financed under retroactive financing*
6. **Summary of the Procurement Packages planned during the first 18 months after project effectiveness ( including those that are subject to retroactive financing and advanced procurement)**

*[List the Packages which require Bank's prior review first and then the other packages]*

Table A.2. Summary of the procurements packages

Ref. No.	Description	Estimated Cost US\$ million	Packages	Domestic Preference (yes/no)	Review by Bank (Prior/Post)	Comments
	Summary of the ICB (Works)	82	5	No	Prior	
	Summary of the ICB (Goods)	43.77	15	No	Prior	
	Summary of the ICB (Works)	64.53	18	No	Post	1st contact for Prior Review
	Summary of the ICB (Goods)	1.86	4	No	Post	1st contact for Prior Review
	Summary of the ICB (Non-Consultant Services)	0.45	1	No	Prior	

**Source** web page *Procurement Plan Template*.

### III. Selection of Consultants

1. **Prior Review Threshold:** Selection decisions subject to Prior Review by Bank as stated in Appendix 1 to the Guidelines Selection and Employment of Consultants:

Table A.3. Selection Method

Ref. No.	Selection Method	Prior Review Threshold	Comments
1.	Competitive Methods (Firms)	Above US\$ 100,000	
2.	Single Source (Firms)	All	
3.	Individual	Above US\$ 100,000	

**Source** web page *Procurement Plan Template*.

2. **Short list comprising entirely of national consultants:** Short list of consultants for services, estimated to cost less than \$300,000 equivalent per contract, may comprise entirely of national consultants in accordance with the provisions of paragraph 2.7 of the Consultant Guidelines.

3. **Any Other Special Selection Arrangements:** *[including advance procurement and retroactive financing, if applicable or delete if not applicable]*

4. **Consultancy Assignments with Selection Methods and Time Schedule**

Table A.4. Consultancy Assignments with Selection Methods and Time Schedule

Ref. No.	Description of Assignment	Estimated Cost US\$ million	Packages	Review by Bank (Prior/Post)	Comments
1.	Summary of the number of contracts that will be lead under QCBC	4.17	7	Prior	
2.	Summary of the number of contracts that will be lead under other methods	0.83	1	Prior	CQS

**Source** web page *Procurement Plan Template*.

---

## Appendix B.

# Software

---

The software in used in this project is free software. All the code was written using the language **R** **Project for Statistical Computing**

“R is available as Free Software under the terms of the Free Software Foundation’s GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.”<sup>1</sup> web page *The R Project for Statistical Computing*,

and *Python*

“Python is developed under an OSI-approved open source license, making it freely usable and distributable, even for commercial use. Python’s license is administered by the Python Software Foundation.” web page *About Python*.

**Python** is general-purpose programming language and **R** is a programming environment optimized for statistics. They’re all widely used in the business and academic worlds, and the modules help users working with those languages to connect to the World Bank Development Indicators API and access the latest data.

To summarize and visualize the data, all the graphs were made with **ggplot2** [Wickham, 2009] a package for **R** written by Hadley Wickham. To create the global map using the spatial data, the project uses **Quantum-Gis (QGIS)**<sup>2 3</sup>. To make the searches at Google using many Amazon Web

---

<sup>1</sup>See Bloomfield, 2014 for an R user guide.

<sup>2</sup>See Cliff and Ord, 1981 for details of what’s spatial data.

<sup>3</sup>GIS refers to Geographical Information Systems and it refers to the set of techniques that uses



Service's computers it uses `Parallel` [Tange, 2011] to distribute the tasks. For the development, edition and writing, this project uses L<sup>A</sup>T<sub>E</sub>X - A document preparation system web page *L<sup>A</sup>T<sub>E</sub>X - A document preparation system*.

Python `pandas` is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

## B.1 Amazon Web Service

Amazon Web Services (AWS) provides computing resources and services that you can use to build applications within minutes at pay-as-you-go pricing. For example, you can rent a server on AWS that you can connect to, configure, secure, and run just as you would a physical server. The difference is the virtual server runs on top of a planet-scale network managed by AWS. The most common service is the Amazon Elastic Compute Cloud (Amazon EC2) machine. EC2 is a web service that provides resizable compute capacity in the cloud. It is designed to make web scale cloud computing easier for developers. Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use. Amazon EC2 provides developers the tools to build failure resilient applications and isolate themselves from common failure scenarios.

The simplest way to connect to an EC2 is by using `ssh` command line tools. You'll specify the private key (.pem) file and `user_name@public_dns_name`.

```
chmod 400 /path/my-key-pair.pem
ssh -i /path/my-key-pair.pem \
ec2-user@ec2-198-51-100-1.compute-1.amazonaws.com
```

To transfer a file to your instance using the instance's public DNS name. For example, if the name of the private key file is `my-key-pair`, the file to transfer is `SampleFile.txt`, and the public DNS name of the instance is

```
ec2-198-51-100-1.compute-1.amazonaws.com
```

use the following command to copy the file to the `ec2-user` home directory.

---

spatial data, see Longley et al., 2005 for more details.

```
scp -i /path/my-key-pair.pem SampleFile.txt \  
ec2-user@ec2-198-51-100-1.compute-1.amazonaws
```

---

## Appendix C.

### World Bank countries

---

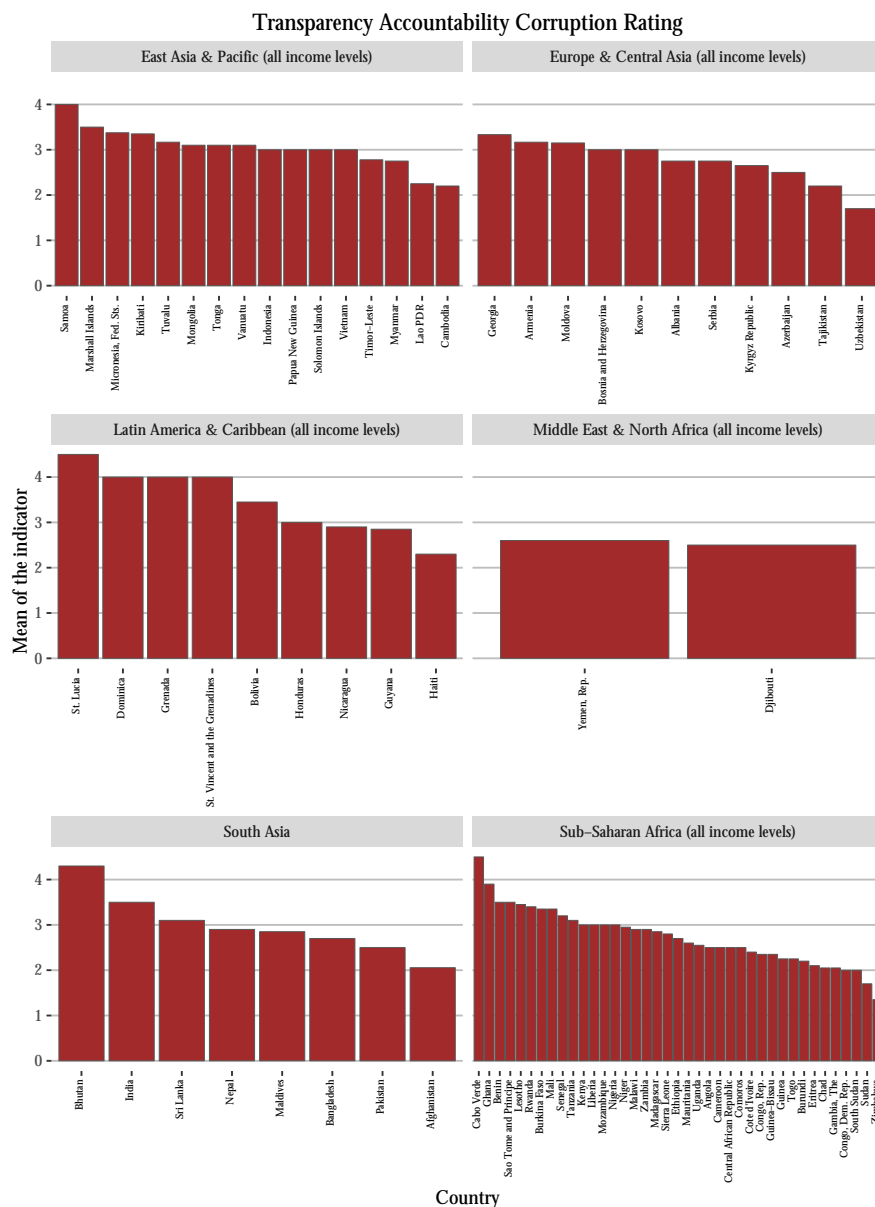
ABW Aruba	BLR Belarus	(IFC classification)
AFG Afghanistan	BLZ Belize	CME Middle East and North Africa
AFR Africa	BMU Bermuda	(IFC classification)
AGO Angola	BOL Bolivia	CMR Cameroon
ALB Albania	BRA Brazil	COD Congo, Dem. Rep.
AND Andorra	BRB Barbados	COG Congo, Rep.
ANR Andean Region	BRN Brunei Darussalam	COL Colombia
ARB Arab World	BTN Bhutan	COM Comoros
ARE United Arab Emirates	BWA Botswana	CPV Cabo Verde
ARG Argentina	CAA Sub-Saharan Africa	CRI Costa Rica
ARM Armenia	(IFC classification)	CSA South Asia
ASM American Samoa	CAF Central African Republic	(IFC classification)
ATG Antigua and Barbuda	CAN Canada	CSS Caribbean small states
AUS Australia	CEA East Asia and the Pacific	CUB Cuba
AUT Austria	(IFC classification)	CUW Curacao
AZE Azerbaijan	CEB Central Europe and the Baltics	CYM Cayman Islands
BDI Burundi	CEU Europe and Central Asia	CYP Cyprus
BEL Belgium	(IFC classification)	CZE Czech Republic
BEN Benin	CHE Switzerland	DEU Germany
BFA Burkina Faso	CHI Channel Islands	DJI Djibouti
BGD Bangladesh	CHL Chile	DMA Dominica
BGR Bulgaria	CHN China	DNK Denmark
BHR Bahrain	CIV Cote d'Ivoire	DOM Dominican Republic
BHS Bahamas, The	CLA Latin America and the	DZA Algeria
BIH Bosnia and Herzegovina	Caribbean	EAP East Asia & Pacific

(developing only)	IMN Isle of Man	(French part)
EAS East Asia & Pacific	IND India	MAR Morocco
(all income levels)	INX Not classified	MCA Central America
ECA Europe & Central Asia	IRL Ireland	MCO Monaco
(developing only)	IRN Iran, Islamic Rep.	MDA Moldova
ECS Europe & Central Asia	IRQ Iraq	MDE Middle East
(all income levels)	ISL Iceland	(developing only)
ECU Ecuador	ISR Israel	MDG Madagascar
EGY Egypt, Arab Rep.	ITA Italy	MDV Maldives
EMU Euro area	JAM Jamaica	MEA Middle East & North Africa
ERI Eritrea	JOR Jordan	(all income levels)
ESP Spain	JPN Japan	MEX Mexico
EST Estonia	KAZ Kazakhstan	MHL Marshall Islands
ETH Ethiopia	KEN Kenya	MIC Middle income
EUU European Union	KGZ Kyrgyz Republic	MKD Macedonia, FYR
FCS Fragile and conflict affected situations	KHM Cambodia	MLI Mali
FIN Finland	KIR Kiribati	MLT Malta
FJI Fiji	KNA St. Kitts and Nevis	MMR Myanmar
FRA France	KOR Korea, Rep.	MNA Middle East & North Africa
FRD Faeroe Islands	KSV Kosovo	(developing only)
FSM Micronesia, Fed. Sts.	KWT Kuwait	MNE Montenegro
GAB Gabon	LAC Latin America & Caribbean	MNG Mongolia
GBR United Kingdom	(developing only)	MNP Northern Mariana Islands
GEO Georgia	LAO Lao PDR	MOZ Mozambique
GHA Ghana	LBN Lebanon	MRT Mauritania
GIN Guinea	LBR Liberia	MUS Mauritius
GMB Gambia, The	LBY Libya	MWI Malawi
GNB Guinea-Bissau	LCA St. Lucia	MYS Malaysia
GNQ Equatorial Guinea	LCN Latin America & Caribbean	NAC North America
GRC Greece	(all income levels)	NAF North Africa
GRD Grenada	LCR Latin America and the Caribbean	NAM Namibia
GRL Greenland	LDC Least developed countries:	NCL New Caledonia
GTM Guatemala	UN classification	NER Niger
GUM Guam	LIC Low income	NGA Nigeria
GUY Guyana	LIE Liechtenstein	NIC Nicaragua
HIC High income	LKA Sri Lanka	NLD Netherlands
HKG Hong Kong SAR, China	LMC Lower middle income	NOC High income: nonOECD
HND Honduras	LMY Low & middle income	NOR Norway
HPC Heavily indebted poor countries (HIPC)	LSO Lesotho	NPL Nepal
HRV Croatia	LTU Lithuania	NZL New Zealand
HTI Haiti	LUX Luxembourg	OECD High income: OECD
HUN Hungary	LVA Latvia	OED OECD members
IDN Indonesia	MAC Macao SAR, China	OMN Oman
	MAF St. Martin	OSS Other small states
		PAK Pakistan

PAN Panama	SOM Somalia	TON Tonga
PER Peru	SRB Serbia	TTO Trinidad and Tobago
PHL Philippines	SSA Sub-Saharan Africa	TUN Tunisia
PLW Palau	(developing only)	TUR Turkey
PNG Papua New Guinea	SSD South Sudan	TUV Tuvalu
POL Poland	SSF Sub-Saharan Africa	TWN Taiwan, China
PRI Puerto Rico	(all income levels)	TZA Tanzania
PRK Korea, Dem. Rep.	SST Small states	UGA Uganda
PRT Portugal	STP Sao Tome and Principe	UKR Ukraine
PRY Paraguay	SUR Suriname	UMC Upper middle income
PSE West Bank and Gaza	SVK Slovak Republic	URY Uruguay
PSS Pacific island small states	SVN Slovenia	USA United States
PYF French Polynesia	SWE Sweden	UZB Uzbekistan
QAT Qatar	SWZ Swaziland	VCT St. Vincent and the Grenadines
ROU Romania	SXM Sint Maarten	VEN Venezuela, RB
RUS Russian Federation	(Dutch part)	VIR Virgin Islands (U.S.)
RWA Rwanda	SXZ Sub-Saharan Africa	VNM Vietnam
SAS South Asia	excluding South Africa	VUT Vanuatu
SAU Saudi Arabia	SYC Seychelles	WLD World
SCE Southern Cone	SYR Syrian Arab Republic	WSM Samoa
SDN Sudan	TCA Turks and Caicos Islands	XZN Sub-Saharan Africa
SEN Senegal	TCD Chad	excluding South Africa and Nigeria
SGP Singapore	TGO Togo	YEM Yemen, Rep.
SLB Solomon Islands	THA Thailand	ZAF South Africa
SLE Sierra Leone	TJK Tajikistan	ZMB Zambia
SLV El Salvador	TKM Turkmenistan	ZWE Zimbabwe
SMR San Marino	TLS Timor-Leste	

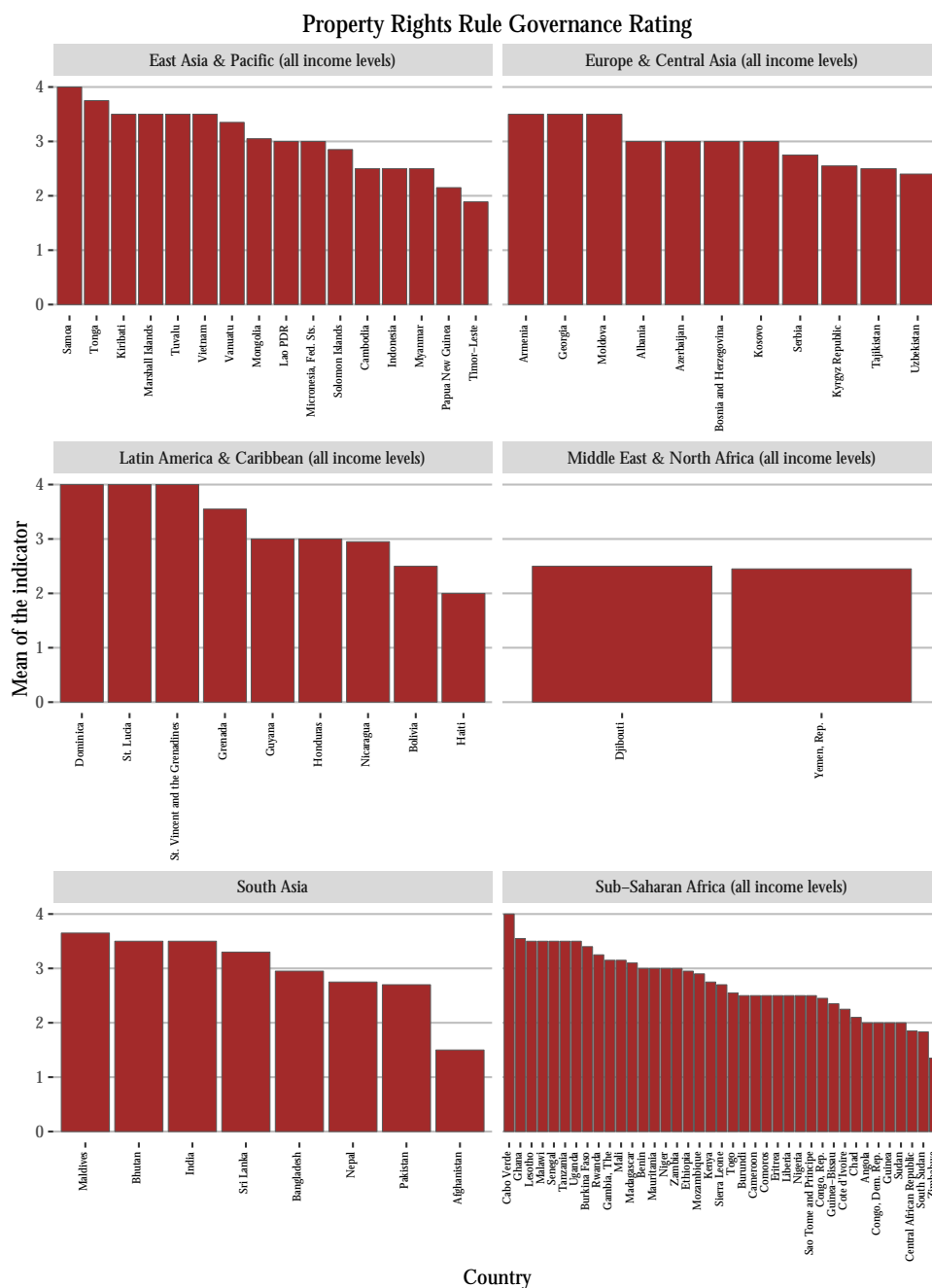
## C.1 World Bank's World Development Indicators

Figure C.1. WDI: Transparency Accountability Corruption Rating, Country/Region



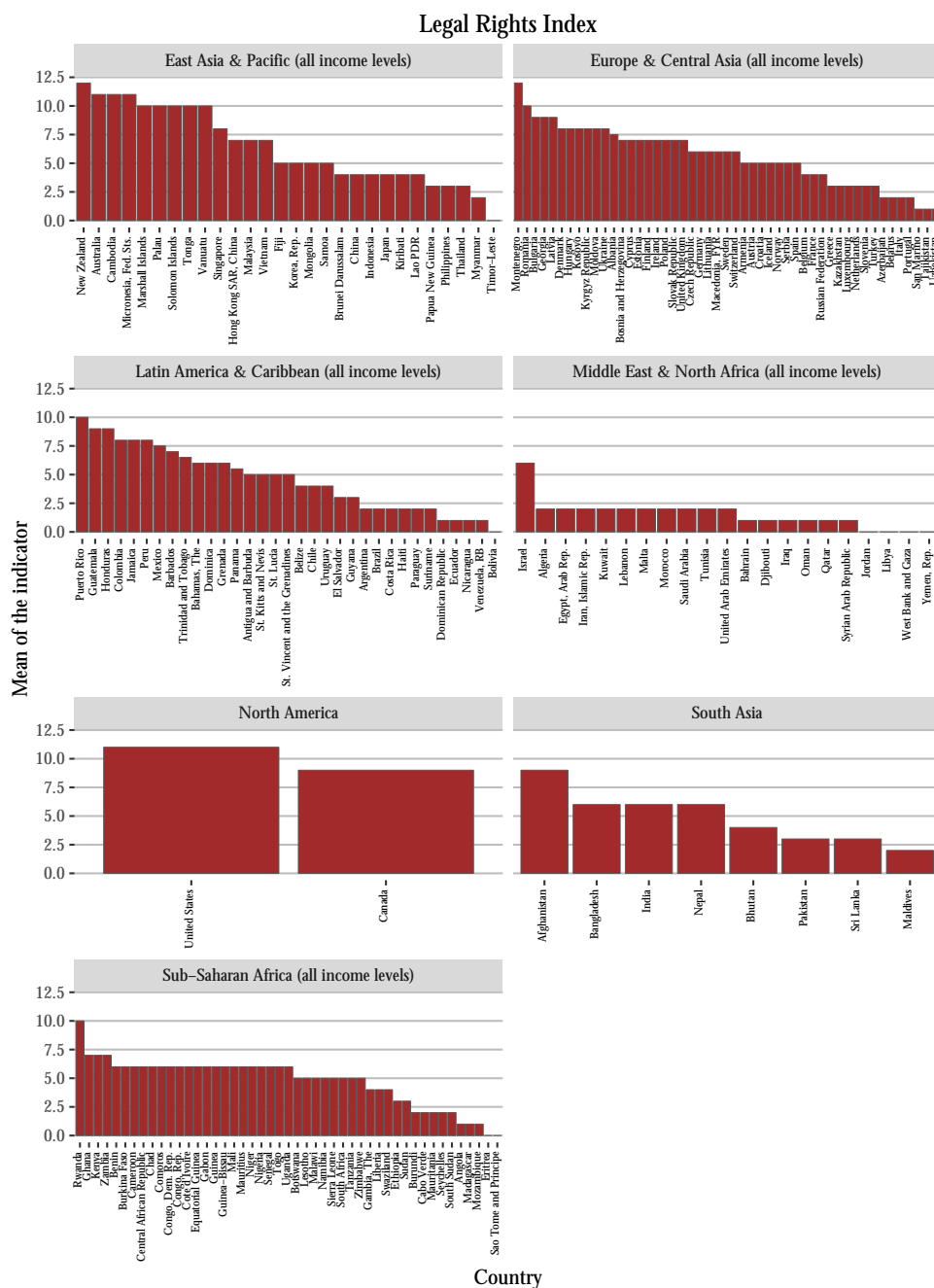
**Source:** Own creation and data obtained with package WDI, see the web page *Package 'WDI'*.

Figure C.2. WDI: Property Rights Rule Governance Rating per Country/Region



**Source:** Own creation and data obtained with package WDI, see the web page *Package 'WDI'*.

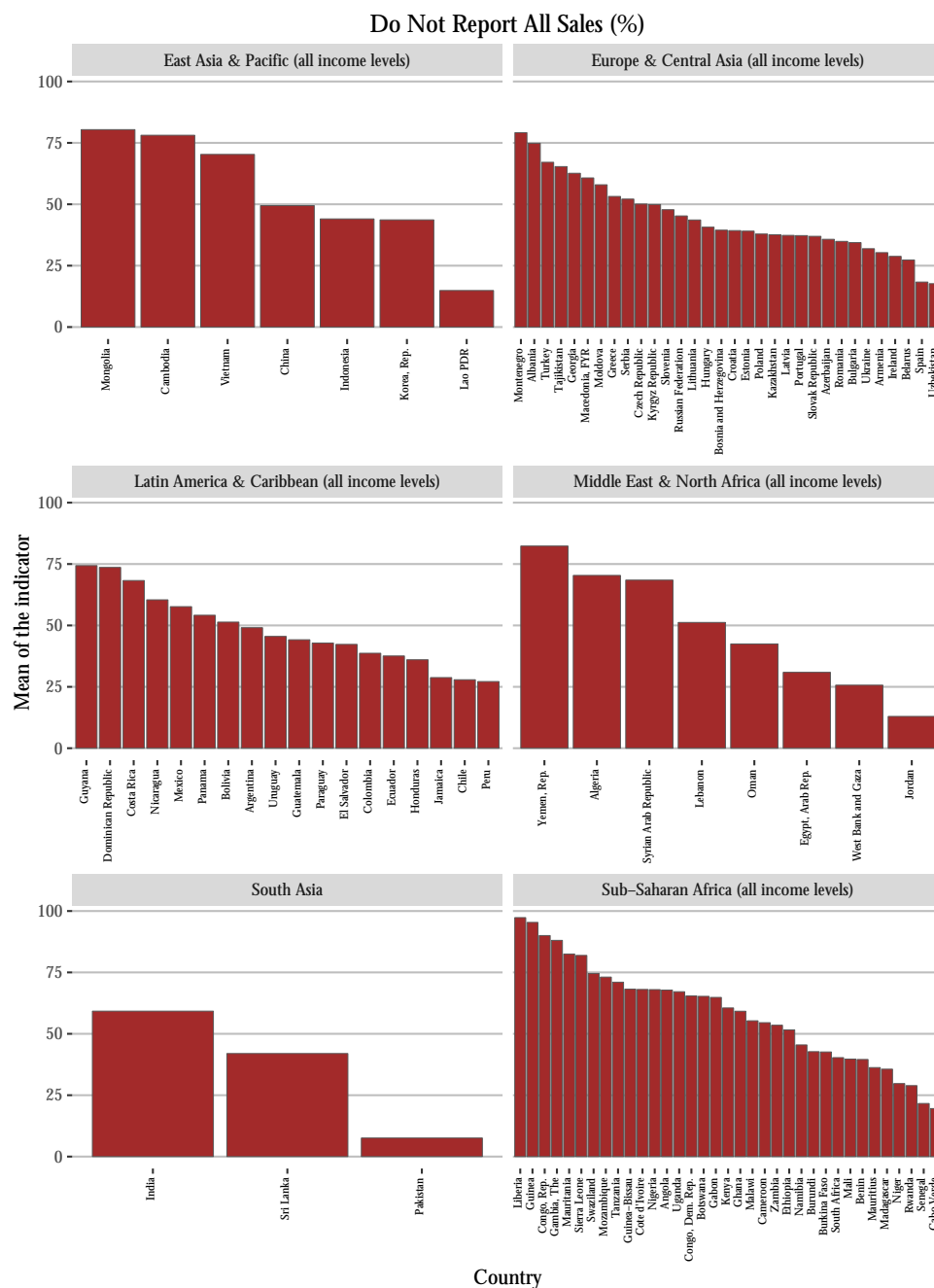
Figure C.3. WDI: Legal Rights Index per Country/Region



**Source:** Own creation and data obtained with package WDI, see the web page *Package 'WDI'*.

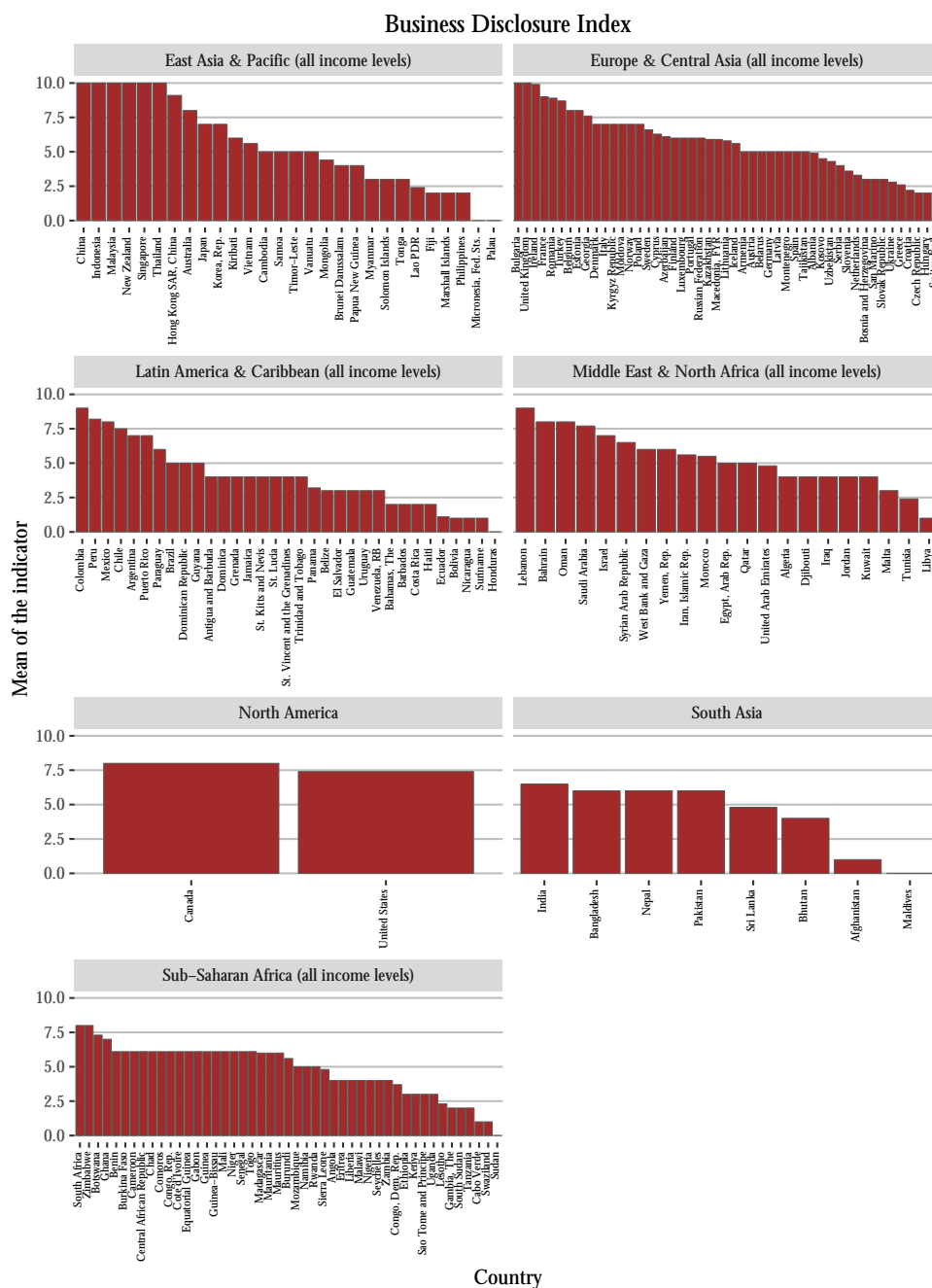


Figure C.4. WDI: % of Firms that do not report all sales per Country/Region



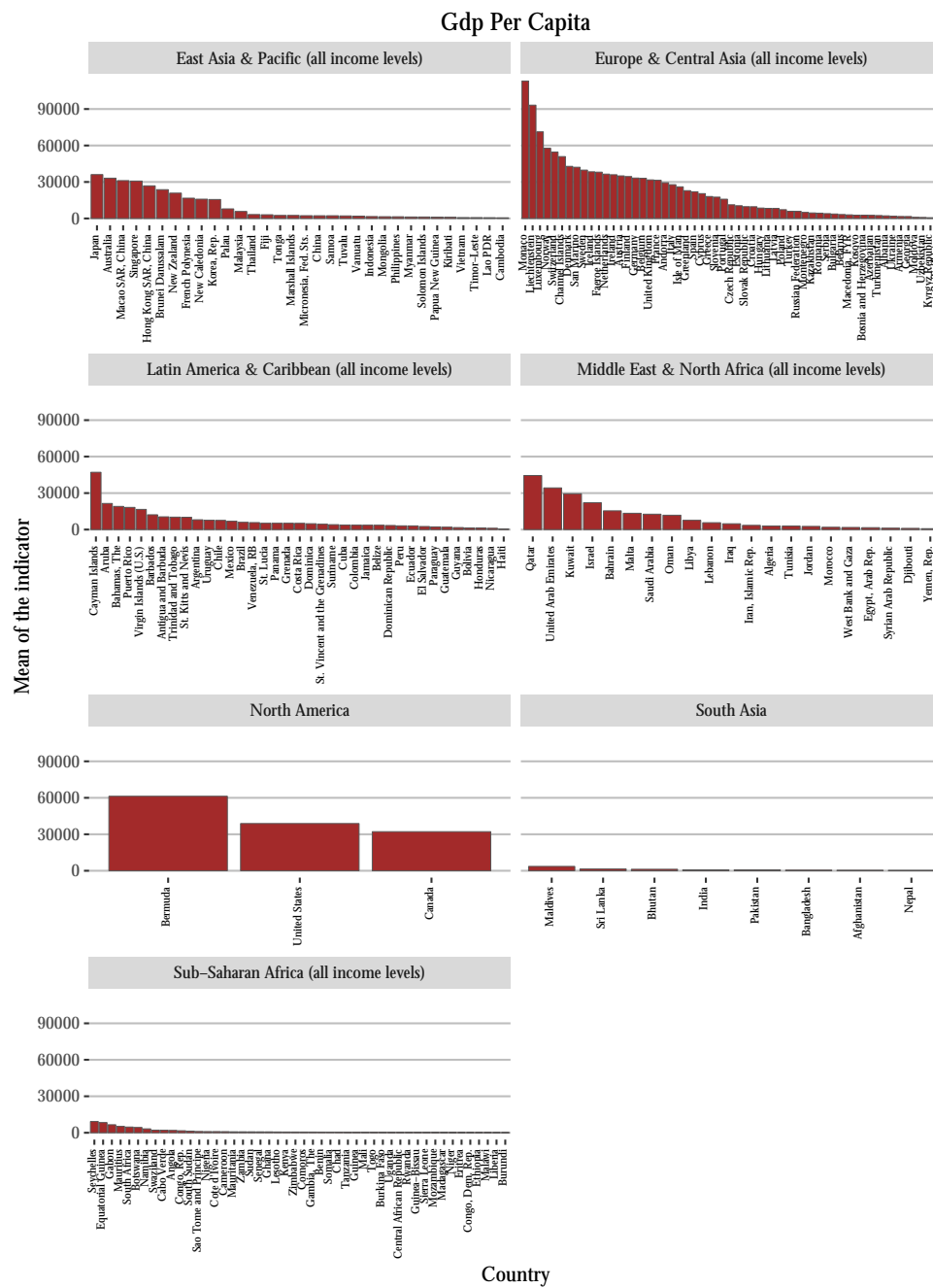
**Source:** Own creation and data obtained with package WDI, see the web page *Package 'WDI'*.

Figure C.5. WDI: Business disclosure index per Country/Region



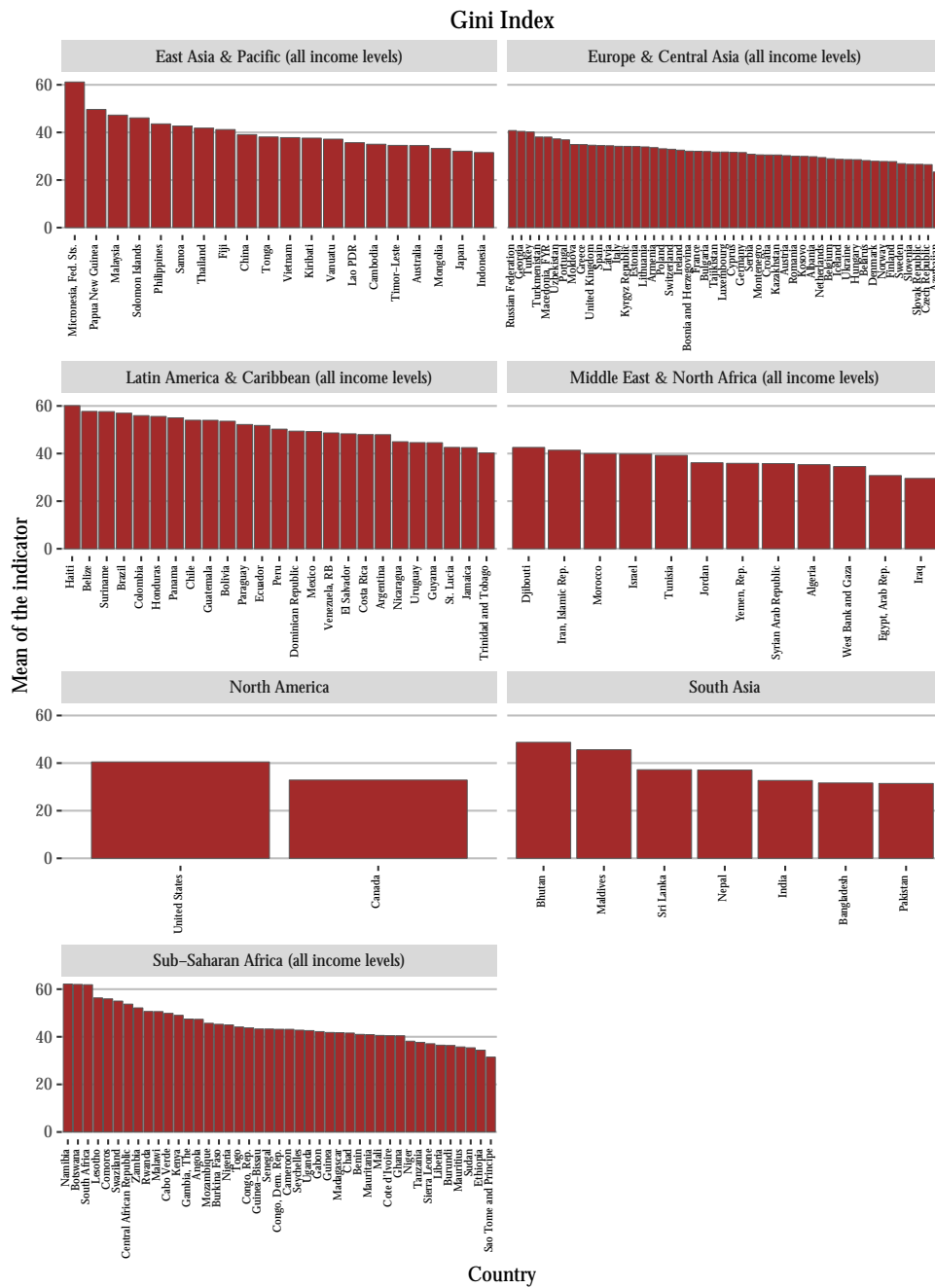
**Source:** Own creation and data obtained with package WDI, see the web page *Package 'WDI'*.

Figure C.6. WDI: Gross Domestic Product per Capita per Country/Region



**Source:** Own creation and data obtained with package WDI, see the web page *Package 'WDI'*.

Figure C.7. WDI: Gini index of inequality per Country/Region



**Source:** Own creation and data obtained with package `WDI`, see the web page *Package ‘WDI’*.

**Primary School Graduation (%)**

**East Asia & Pacific (all income levels)**

Country	Graduation (%)
Japan	100
Hong Kong SAR, China	100
Korea, Rep.	100
Singapore	98
Malaysia	95
Brunei Darussalam	92
Vietnam	90
Taiwan	88
Fiji	85
Thailand	85
Indonesia	85
Sri Lanka	82
China	80
Mongolia	78
Macao SAR, China	75
Marshall Islands	75
Kiribati	72
Timor-Leste	72
Philippines	68
Vanuatu	65
Myanmar	62
Solomon Islands	58
Lao PDR	55
Cambodia	50
Papua New Guinea	48

**Europe & Central Asia (all income levels)**

Country	Graduation (%)
Croatia	100
Slovenia	100
Czech Republic	100
Finland	100
Netherlands	100
Sweden	100
Slovakia	100
Kazakhstan	100
Ukraine	100
Slovak Republic	100
Austria	100
Latvia	100
Uzbekistan	100
Lithuania	100
Spain	100
Belgium	100
France	100
Germany	100
Poland	100
Italy	100
Turkey	100
Armenia	100
Kyrgyz Republic	100
Georgia	100
Russian Federation	100
San Marino	100
Cyprus	100
Bulgaria	100
Belgium	100
Bosnia and Herzegovina	95
Montenegro	90
Algeria	85
Switzerland	80

**Latin America & Caribbean (all income levels)**

Country	Graduation (%)
Cuba	95
Aruba	95
Barbados	95
Suriname	95
Costa Rica	95
Antigua and Barbuda	95
Uruguay	95
Argentina	95
Trinidad and Tobago	95
Mexico	95
Jamaica	95
Dominica	95
Bahamas, The	95
Guatemala	95
Cayman Islands	95
Costa Rica	95
Peru	95
Belize	95
Suriname	95
Colombia	95
St. Kitts and Nevis	95
Ecuador	95
Brazil	95
Paraguay	95
Colombia	95
St. Vincent and the Grenadines	95
Dominican Republic	95
El Salvador	95
Honduras	95
Nicaragua	95
Haiti	95

**Middle East & North Africa (all income levels)**

Country	Graduation (%)
Israel	100
Egypt, Arab Rep.	100
Saudi Arabia	100
Malta	100
West Bank and Gaza	100
Jordan	100
Kuwait	100
Lebanon	100
Bahrain	100
Algeria	100
Iran, Islamic Rep.	100
Tunisia	100
Syrian Arab Republic	100
Qatar	100
Oman	100
United Arab Emirates	100
Djibouti	100
Morocco	100
Yemen, Rep.	100
Iraq	100

**North America**

Country	Graduation (%)
Canada	100
Bermuda	95

**South Asia**

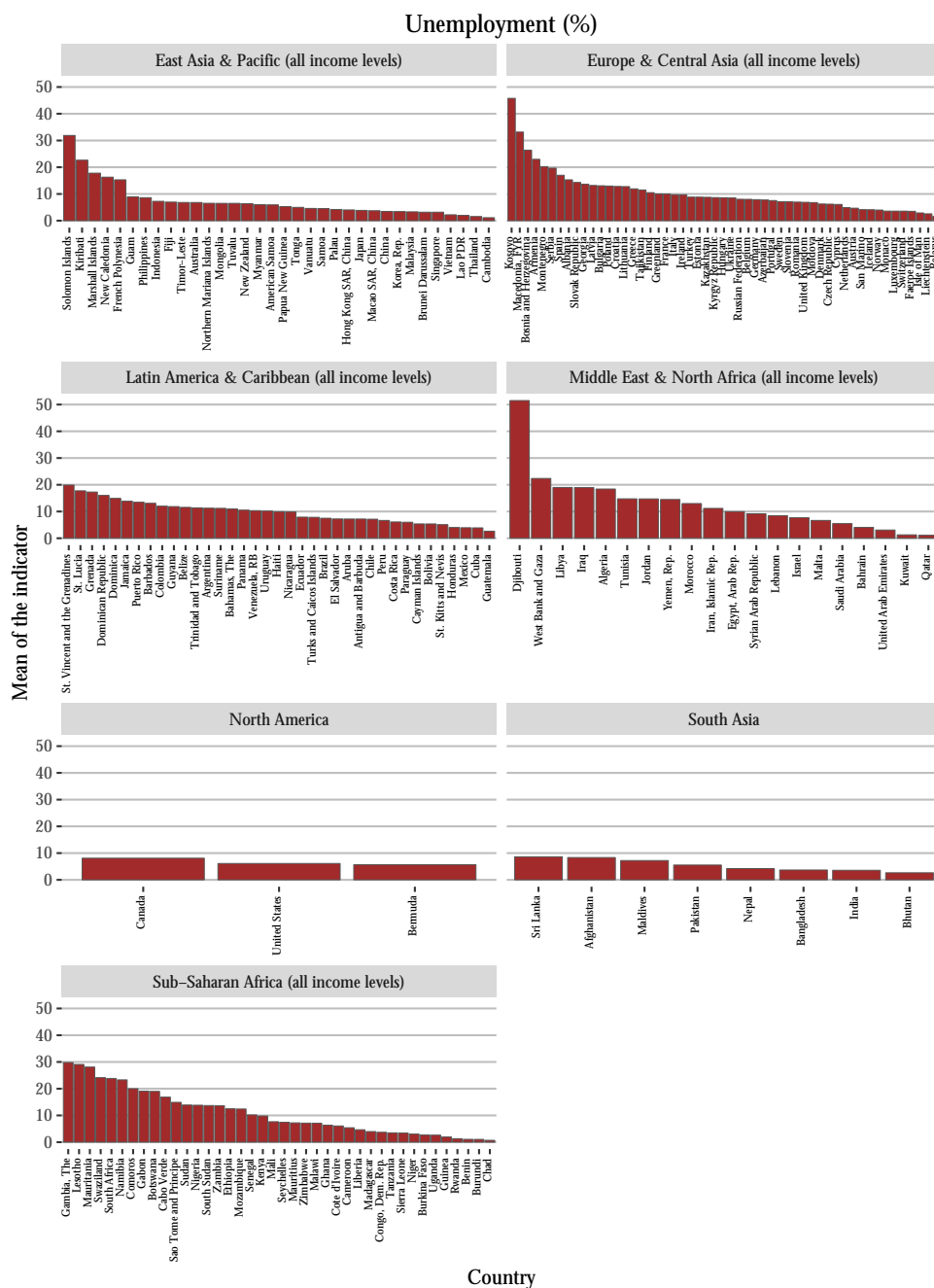
Country	Graduation (%)
Sri Lanka	100
Afghanistan	95
Maldives	95
Bhutan	95
Bangladesh	95
Pakistan	95
India	95
Nepal	95

**Sub-Saharan Africa (all income levels)**

Country	Graduation (%)
Mauritius	95
Swaziland	95
Botswana	95
Kenya	95
Namibia	95
Tanzania	95
Eritrea	95
South Africa	95
Nigeria	95
Cote d'Ivoire	95
Senegal	95
Burkina Faso	95
Sierra Leone	95
Gambia, The	95
Zambia	95
Lesotho	95
Benin	95
Congo, Rep.	95
Congo, Dem. Rep.	95
Equatorial Guinea	95
Sierra Leone	95
Gabon	95
Comoros	95
Madagascar	95
Ethiopia	95
Kenya	95
Rwanda	95
Malawi	95
Chad	95
Angola	95
Mozambique	95
Uganda	95

74

Figure C.9. WDI: % of Unemployment per Country/Region



**Source:** Own creation and data obtained with package WDI, see the web page *Package 'WDI'*.

---

## Appendix D.

### Code

---

*Code D.1.* WDI Download code

```
library(WDI); library(plyr); library(dplyr); library(tidyr)

countries <- read.csv('../data_clean/clean_country_names.csv') %>%
  dplyr::rename(country=dirty, country_clean=clean)

#List of the indicators for the project.
indicators <- c('IC.BUS.DISC.XQ', 'IC.FRM.CMPU.ZS',
  'IC.FRM.CORR.ZS', 'IC.FRM.INFM.ZS',
  'IC.LGL.CRED.XQ', 'IC.LGL.DURS', 'IC.TAX.GIFT.ZS',
  'IQ.CPA.PROP.XQ', 'IQ.CPA.TRAN.XQ', 'NY.GDP.PCAP.CD',
  'SE.PRM.PRSL.ZS', 'SI.POV.GINI', 'SL.UEM.TOTL.NE.ZS')

indicators_names <- c('business_disclosure_index',
  'firms_competing_against_informal_firms_perc',
  'payments_to_public_officials_perc',
  'do_not_report_all_sales_perc', 'legal_rights_index',
  'time_to_enforce_contract', 'bribes_to_tax_officials_perc',
  'property_rights_rule_governance_rating',
  'transparency_accountability_corruption_rating',
```

```
'gdp_per_capita', 'primary_school_graduation_perc',
'gini_index', 'unemployment_perc')

df <- WDICache()

df_indicators <- ldply(indicators, function(x){
  data.frame(t(
    WDIsearch(string=x,field=c('indicator','name'), cache = df)
  )))

df_indicators$indicator_name <- indicators_names

df_wdi <- WDI(country = 'all', indicator = df_indicators$indicator,
  start = 1990, end=2014,extra=T, cache=df ) %>%
  gather(indicator, value,
    IC.BUS.DISC.XQ:SL.UEM.TOTL.NE.ZS) %>%
  filter(!is.na(value)) %>%
  filter(region!='Aggregates')

df_wdi <- left_join(df_wdi,
  select(df_indicators,indicator,indicator_name))

df_wdi_countries <- df_wdi %>%
  filter(country %in% countries$country) %>%
  left_join(countries)

write.csv(df_wdi, file = '../data_clean/WDI.csv',row.names =
  F,quote =T)
```



## D.1 Entity Names disambiguation code

*Code D.2. Searches for canonical names*

```
# -*- coding: utf-8 -*-
# Setup the AWS Code
machine_image='ami-f44a869c'
instance_type='m1.small'
key_name='carpetri'
key_extension='.pem'
key_dir='~/.ssh'
login_user='ec2-user'
group_name='worldbank'
max_instance_count = 450
query_block_size=100

import os
import time
import boto
import boto.manage.cmdshell
import pandas as pd
from time import sleep

def launch_instance(ami='ami-7341831a',
                    instance_type='t1.micro',
                    key_name='jeff',
                    key_extension='.pem',
                    key_dir='~/.ssh',
                    group_name='worldbank',
                    ssh_port=22,
                    cidr='0.0.0.0/0',
                    tag='paws',
                    user_data=None,
                    cmd_shell=True,
                    login_user='ec2-user',
```

```
        ssh_passwd=None,
        wait_to_run=False):
    """
    Launch an instance and wait for it to start running.
    Returns a tuple consisting of the Instance object and the
        CmdShell
    object, if request, or None.

    ami          The ID of the Amazon Machine Image that this
        instance will
                be based on.  Default is a 64-bit Amazon Linux EBS
                image.

    instance_type The type of the instance.

    key_name      The name of the SSH Key used for logging into the
        instance.
                It will be created if it does not exist.

    key_extension The file extension for SSH private key files.

    key_dir       The path to the directory containing SSH private
        keys.
                This is usually ~/.ssh.

    group_name    The name of the security group used to control access
        to the instance.  It will be created if it does not
        exist.

    ssh_port      The port number you want to use for SSH access
        (default 22).

    cidr          The CIDR block used to limit access to your instance.

    tag          A name that will be used to tag the instance so we
```

```
can
    easily find it later.

user_data Data that will be passed to the newly started
    instance at launch and will be accessible via
    the metadata service running at
    http://169.254.169.254.

cmd_shell If true, a boto CmdShell object will be created and
    returned.
    This allows programmatic SSH access to the new
    instance.

login_user The user name used when SSH'ing into new instance.
    The
    default is 'ec2-user'

ssh_passwd The password for your SSH key if it is encrypted
    with a
    passphrase.
"""
cmd = None

# Create a connection to EC2 service.
# You can pass credentials in to the connect_ec2 method
    explicitly
# or you can use the default credentials in your ~/.boto config
    file
# as we are doing here.
ec2 = boto.connect_ec2()

# Check to see if specified keypair already exists.
# If we get an InvalidKeyPair.NotFound error back from EC2,
# it means that it doesn't exist and we need to create it.
try:
```

```
key = ec2.get_all_key_pairs(keynames=[key_name])[0]
except ec2.ResponseError, e:
    if e.code == 'InvalidKeyPair.NotFound':
        print 'Creating keypair: %s' % key_name
        # Create an SSH key to use when logging into instances.
        key = ec2.create_key_pair(key_name)

        # AWS will store the public key but the private key is
        # generated and returned and needs to be stored locally.
        # The save method will also chmod the file to protect
        # your private key.
        key.save(key_dir)
    else:
        raise

# Check to see if specified security group already exists.
# If we get an InvalidGroup.NotFound error back from EC2,
# it means that it doesn't exist and we need to create it.
try:
    group =
        ec2.get_all_security_groups(groupnames=[group_name])[0]
except ec2.ResponseError, e:
    if e.code == 'InvalidGroup.NotFound':
        print 'Creating Security Group: %s' % group_name
        # Create a security group to control access to instance
        # via SSH.
        group = ec2.create_security_group(group_name,
                                           'A group that allows
                                           SSH access')
    else:
        raise

# Add a rule to the security group to authorize SSH traffic
# on the specified port.
try:
```

```
group.authorize('tcp', ssh_port, ssh_port, cidr)
except ec2.ResponseError, e:
    if e.code == 'InvalidPermission.Duplicate':
        print 'Security Group: %s already authorized' %
            group_name
    else:
        raise

# Now start up the instance. The run_instances method
# has many, many parameters but these are all we need
# for now.
reservation = ec2.run_instances(ami,
                                key_name=key_name,
                                security_groups=[group_name],
                                instance_type=instance_type,
                                user_data=user_data)

# Find the actual Instance object inside the Reservation object
# returned by EC2.

instance = reservation.instances[0]

# The instance has been launched but it's not yet up and
# running. Let's wait for it's state to change to 'running'.

if wait_to_run:
    print 'waiting for instance'
    while instance.state != 'running':
        print '.'
        time.sleep(5)
        instance.update()
    print 'done'

# Let's tag the instance with the specified label so we can
# identify it later.
```

```

instance.add_tag(tag)
# The instance is now running, let's try to programmatically
# SSH to the instance using Paramiko via boto CmdShell.
#if cmd_shell:
#    key_path = os.path.join(os.path.expanduser(key_dir),
#                             key_name+key_extension)
#    cmd =
#        boto.manage.cmdshell.sshclient_from_instance(instance,
#                                                       key_path,
#                                                       user_name=login_user)
#return (instance, cmd)
return instance

# Define the Calculations to Be Run on Each Instance
startup_script = """#!/home/$(login_user)s/anaconda/bin/python
##### Get Data
print("Getting Data")
entity_block_start = %(entity_block_start)i
entity_block_stop = %(entity_block_stop)i #Inclusive
import pandas as pd
from pylab import *

entity_names = pd.DataFrame(%(entity_names)r,
                             columns=["Name"],
                             index=%(index)r)
#pd.read_csv('../Data/Entities/all_entities.csv',index_col=0)

#### Establish parameters
error_method = "sleep"
max_attempt_count = 10

import sys
sys.path.append('/home/$(login_user)s/')
import google

```

```
from urllib2 import HTTPError

def perform_query(entity_name):
    query = google.search(entity_name, stop=1, only_standard=True,
        pause=0)
    return [url for url in query]

##### Build system for handling errors.
def handle_error(method="wait", sleep_time=5):
    if method=='wait':
        print "Waiting and trying again."
        sleep(sleep_time)
    elif method=='tor':
        print "Resetting Tor node and cookies."
        reset_tor()

if error_method=='Tor': #NEEDS ROOT ACCESS TO RESET!
    import urllib2
    proxy = urllib2.ProxyHandler({'http': '127.0.0.1:8118'})
    opener = urllib2.build_opener(proxy)
    urllib2.install_opener(opener)
    from os import system
    #print urllib2.urlopen('http://icanhazip.com/').read()

    import subprocess

    def reset_tor():
        proc = subprocess.Popen(["ps aux | grep tor | grep -v
            grep"], stdout=subprocess.PIPE, shell=True)
        (out, err) = proc.communicate()
        target_process = out.split()[1]
        system("sudo kill -s SIGHUP %s"%%target_process) #NEEDS
            ROOT ACCESS!

    reset_tor()
```

```
elif error_method=='sleep':
    from time import sleep

#### Initialize Data
query_results =
    pd.DataFrame(index=arange(entity_block_start,entity_block_stop+1),
        columns=["Name"]+map(str, range(10)))

#### Run queries
print("Running Queries")
for i in arange(entity_block_start,entity_block_stop+1):
    entity_name = entity_names.ix[i,"Name"]
    print i, entity_name

    attempt_count = 0
    while attempt_count<max_attempt_count:
        try:
            urls = perform_query(entity_name)
            break
        except HTTPError:
            print "HTTP Error."
            handle_error(method=error_method)
            google.cookie_jar.clear()
            attempt_count += 1
    else:
        raise Exception("Failed after %%i attempts to query entry
            %%i, name %%"%(attempt_count,i,entity_name))

    query_results.ix[i,"Name"] = entity_name
    n_urls = min(10,len(urls))
    query_results.ix[i,map(str, range(n_urls))] = urls[:n_urls]
```



```

#### Save results
print("Saving Results")
import csv
query_results.to_csv(
    "/home/%(login_user)s/all_entities_Google_results_
    %(entity_block_start)i_to_%(entity_block_stop)i.csv",
    quoting=csv.QUOTE_ALL)
query_results.to_hdf(
    "/home/%(login_user)s/all_entities_Google_results_
    %(entity_block_start)i_to_%(entity_block_stop)i.h5", 'df')
"""
# Load Entity Names and Define Blocks

entity_names =
    pd.read_csv('../Data/Entities/all_entities.csv', index_col=0)
entity_block_starts =
    entity_names.index[::query_block_size].tolist()
entity_block_stops =
    entity_names.index[query_block_size-1::query_block_size].tolist()
    + [entity_names.index[-1]]

# Create Query Commands and Send Out to New Instances
print("Sending Out Queries to New Instances")
from os import listdir
dirlist = listdir("../Search_Results/")
instances = []
instance_count = 0
for start, stop in zip(entity_block_starts, entity_block_stops):
    print start, stop

    if "all_entities_Google_results_%i_to_%i.csv"%(start, stop) in
        dirlist:
        print("Data already exists. Skipping.")
        continue #We've already done this condition, so don't
            requery it

```

```
entities = entity_names.ix[start:stop, "Name"]
parameters = {
    'index': entities.index.values,
    'entity_names': entities.values,
    'entity_block_start': start,
    'entity_block_stop': stop,
    'login_user': login_user
}

try:
    instance =
        launch_instance(user_data=startup_script%parameters,
                        ami=machine_image,
                        instance_type=instance_type,
                        key_name=key_name,
                        key_extension=key_extension,
                        key_dir=key_extension,
                        group_name=group_name,
                        tag='world_bank_query')

    instance.add_tag("query_block_start", value=start)
    instance.add_tag("query_block_stop", value=stop)
    instances.append((instance))
    instance_count+=1
except:
    pass
if instance_count>max_instance_count:
    break

# Wait for Calculations to Run
print("Waiting for Calculations to Run")
sleep(900) #15 minutes

# Pull in Data and Terminate Instances
```

```
print("Pulling in Data and Terminating Instances")
key_path = os.path.join(os.path.expanduser(key_dir),
                        key_name+key_extension)

make_cmd = lambda
    instance:boto.manage.cmdshell.sshclient_from_instance(
        instance,key_path,user_name=login_user)

for instance in instances:
    print instance
    instance.update()
    if instance.state in ['shutting-down', 'terminated', 'pending']:
        print("Instance state is %, skipping"%instance.state)
        continue
    else:
        start = int(instance.tags['query_block_start'])
        stop = int(instance.tags['query_block_stop'])
        print("Pulling query block %i to %i"%(start,stop))
        try:
            cmd = make_cmd(instance)
            cmd.get_file("all_entities_Google_results_%i_to_%i.csv"
                        % (start,stop),
                        "../Search_Results/all_entities_Google_results_
                        %i_to_%i.csv" % (start,stop))
            cmd.get_file("all_entities_Google_results_%i_to_%i.h5"
                        % (start,stop),
                        "../Search_Results/all_entities_Google_results_
                        %i_to_%i.h5" % (start,stop))
        except (IOError, AttributeError):
            print("Data not calculated for this instance.")
            pass
    instance.terminate()
```

*Code D.3. Cluster canonical names*

```

# -*- coding: utf-8 -*-
import pandas as pd
from numpy import intersect1d
from scipy import sparse
import networkx as nx
import csv

# Load Data, Sort It, and Drop Header Rows from CSV File Format
queries =
    pd.read_csv('../Search_Results/all_entities_Google_results.csv',
        index_col=0)
queries = queries.sort_index()
queries =
    queries.drop(queries.index[where(queries["Name"]=="Name")[0]])
queries.to_csv('../Search_Results/all_entities_Google_results.csv',
    quoting=csv.QUOTE_ALL)
query_results = queries.values[:,1:]

# Define Comparison and Linking Function
def compare_ij(i,j):
    return [y for y in ##All the elements
        set(query_results[i]).intersection(query_results[j])
        ###That are in both sets of query results
        if type(y)==str] ###That are strings (i.e. Not nans)
#     Use the below code if you wanted to keep track of rankings.
#     It's slower by up to a factor of 3 in the worst case.
#     dict1 = {}
#     dict2 = {}
#     list1 = query_results[i]
#     list2 = query_results[j]
#     indices = xrange(10)
#     for ele in indices:
#         dict1[list1[ele]] = ele
#     for ele in indices:

```

```
#         value = list2[ele]
#         if value in dict1:
#             dict2[value] = ele
#     intersect = [y for y in dict2.keys() if type(y)==str]
#     ranks = [(dict2[value],dict1[value]) for value in intersect]
#     return intersect, ranks

n_entities = len(query_results)
import codecs
generate_edgelist = nx.readwrite.edgelist.generate_edgelist
delimiter=";;;"

def clean_string_ends(s,forbidden_character=','):
    while s.startswith(','):
        s = s[1:]
    while s.endswith(','):
        s = s[:-1]
    return s

def find_links(i):
    match_graph = nx.Graph()
    entity_i = queries.ix[i,"Name"]
    print i, entity_i

    entity_i = clean_string_ends(entity_i)

    for j in arange(i+1, n_entities):
        entity_j = queries.ix[j,"Name"]

        entity_j = clean_string_ends(entity_j)

        matches = compare_ij(i,j)
        attr_dict = {"matches": matches,
                    "n_matches": len(matches),
                    }
```

```
        if matches:
            match_graph.add_edge(entity_i, entity_j,
                                  attr_dict=attr_dict)
    if match_graph.number_of_nodes():
        fh=codecs.open( '../Search_Clustering/%i_matches.el' %i,
                        mode='w', encoding='utf-8')
        for line in generate_edgelist(match_graph, delimiter,
                                      data=True):
            line+='\n'
            fh.write(line)
        fh.close()

from multiprocessing import Pool
p = Pool()

%%prun
p.map(find_links, arange(n_entities))
```

# Bibliography

---

## Books

Amazon Web Service (2015). *Getting Started with AWS*. Amazon Web Services, Inc.  
URL: <http://s3.amazonaws.com/awsdocs/gettingstarted/latest/awsgsg-intro.pdf>.

Bloomfield, Victor A. (2014). *Using R for Numerical Analysis in Science and Engineering*. Chapman & Hall/CRC. ISBN: 978-1439884485. URL: <http://www.crcpress.com/product/isbn/9781439884485>.

Cliff, Andrew D. and J. K. Ord (1981). *Spatial Processes: Models & Applications*. Pion Limited.

Longley, Paul A. et al. (2005). *Geographical Information Systems: Principles, Techniques, Applications and Management*. 2nd ed. Abridged.

Wickham, Hadley (2009). *ggplot2: Elegant graphics for data analysis*. Springer. URL: <http://had.co.nz/ggplot2/book>.

## Articles

Gagnon, Francis and Betsy Wiramidjaja (2014). “Poster: What does Corruption look like?” *Integrity Vice Presidency, The World Bank. Washington, D.C.*

- Tange, O. (2011). “GNU Parallel - The Command Line Power Tool”. *The USENIX Magazine* 36.num. 1, 42 a 47. URL: <http://www.gnu.org/s/parallel>.
- The World Bank (2011a). “Guidelines: Procurement of goods, works, and non-consulting services under IBRD loans and IDA credits and grants”. *The International Bank for Reconstruction and Development, Washington, D.C.*
- (2011b). “Guidelines: Selection and Employment of Consultants under IBRD Loans and IDA Credits and Grants”. *The International Bank for Reconstruction and Development, Washington, D.C.*

## Other references

- Amazon Web Service (2014). *Amazon EC2 - Virtual Server Hosting*. [https://aws.amazon.com/ec2/?nc1=f\\_ls](https://aws.amazon.com/ec2/?nc1=f_ls). (Visited on 12/01/2015).
- Artem Zubkov (2013). *World Bank Contract Awards*. <http://artzub.com>. (Visited on 12/01/2015).
- Data Sience for Social Good, University of Chicago (2014). *Clean Development: Data Mining for Corruption Risks*. <http://dssg.uchicago.edu/2014/07/11/clean-development-data-mining-for-corruption-risks/>. (Visited on 12/01/2015).
- (2015). *What Makes a Good DSSG Project?* <http://dssg.uchicago.edu/2015/11/04/what-makes-a-good-dssg-project/>. (Visited on 12/01/2015).
- GNU project. *L<sup>A</sup>T<sub>E</sub>X - A document preparation system*. <http://www.latex-project.org>. (Visited on 12/01/2015).
- *The R Project for Statistical Computing*. <http://www.r-project.org/about.html>. (Visited on 12/01/2015).
- Oliver Sherouse (2014). *Welcome to wldata’s documentation*. <https://wldata.readthedocs.org/en/latest/>. (Visited on 12/01/2015).



- Python Software Foundation. *About Python*. <https://www.python.org/about/>. (Visited on 12/01/2015).
- The World Bank (2011c). *Investigations*. <http://www.worldbank.org/en/about/unit/integrity-vice-presidency/what-is-fraud-and-corruption>. (Visited on 12/01/2015).
- (2011d). *Procurement Plan Template*. [http://siteresources.worldbank.org/INTPROCUREMENT/Resources/Sample\\_of\\_Summarized\\_Procurement\\_Plan.docx](http://siteresources.worldbank.org/INTPROCUREMENT/Resources/Sample_of_Summarized_Procurement_Plan.docx). (Visited on 12/01/2015).
- (2011e). *What is Fraud and Corruption?* <http://www.worldbank.org/en/about/unit/integrity-vice-presidency/what-is-fraud-and-corruption>. (Visited on 12/01/2015).
- (2011f). *World Bank Listing of Ineligible Firms and Individuals*. <http://web.worldbank.org/external/default/main?contentMDK=64069844&menuPK=116730&pagePK=64148989&piPK=64148984&querycontentMDK=64069700&theSitePK=84266>. (Visited on 12/01/2015).
- (2014a). *Accessing the World Bank Data APIs in Python, R, Ruby and Stata*. <http://blogs.worldbank.org/opendata/accessing-world-bank-data-apis-python-r-ruby-stata>. (Visited on 12/01/2015).
- (2014b). *It Takes Villages to Conserve Indonesia's Precious Coral Reefs*. <http://www.worldbank.org/en/news/feature/2014/06/05/it-takes-villages-to-serve-indonesia-precious-coral-reefs>. (Visited on 12/01/2015).
- (2014c). *New roads and irrigation systems improve life in Ecuador*. <http://www.worldbank.org/en/news/feature/2014/06/04/nuevas-carreteras-y-sistemas-de-riego-mejoran-la-vida-en-ecuador>. (Visited on 12/01/2015).
- (2014d). *Transformational Hydropower Development Project Paves the Way for 9 Million People in the Democratic Republic of Congo to Gain Access to Elec-*

*tricity*. <http://www.worldbank.org/en/news/feature/2014/03/20/transformational-hydropower-development-project-paves-the-way-for-9-million-people-in-the-democratic-republic-of-congo-to-gain-access-to-electricity>. (Visited on 12/01/2015).

The World Bank (2015a). *Data Bank*. <http://data.worldbank.org>. (Visited on 12/01/2015).

— (2015b). *History of the World Bank*. <http://www.worldbank.org/en/about/history>. (Visited on 12/01/2015).

— (2015c). *Integrity Complaint Form*. [https://intlbankforreconanddev.ethicspointvp.com/custom/ibrd/\\_crf/english/form\\_data.asp](https://intlbankforreconanddev.ethicspointvp.com/custom/ibrd/_crf/english/form_data.asp). (Visited on 12/01/2015).

— (2015d). *Procurements Rules*. <http://go.worldbank.org/ZYEAU7SVT0>. (Visited on 12/01/2015).

— (2015e). *What we do*. <http://www.worldbank.org/en/about/what-we-do>. (Visited on 12/01/2015).

Vincent Arel-Bundock (2014). *Package ‘WDI’*. <https://cran.r-project.org/web/packages/WDI/WDI.pdf>. (Visited on 12/01/2015).