Mining Large-scale Corporate Networks

Frank Takes

LIACS, Leiden University
AISSR, University of Amsterdam

Leiden Complex Networks Network — LCN2
January 29, 2016
Introduction
Introduction
Introduction

Introduction


Leiden Institute of Advanced Computer Science (LIACS)

Amsterdam Institute for Social Science Research (AISSR)
Introduction

- Leiden Institute of Advanced Computer Science (LIACS)
- Amsterdam Institute for Social Science Research (AISSR)
- Interdisciplinary research
Corporate networks

- Networks
- Nodes are firms
- Edges/links indicate for example:
  - trading
  - borrowing/lending
  - ownership
  - (board) interlocks
- Aim is to understand:
  - Corporate control
  - Economy at a macro level
  - Corporate elites
Corporate network

Figure: Board interlock network of 30 firms.
Board interlocks

- **Board interlock**: there is a relationship between firms because they share a board member or director
Board interlocks

- **Board interlock**: there is a relationship between firms because they share a board member or director
- “... a personal union, so to speak, is established between the banks and the biggest industrial and commercial enterprises, the merging of one with another through the acquisition of shares, through the appointment of bank directors to the Supervisory Boards (or Boards of Directors) of industrial and commercial enterprises, and vice versa.”
Board interlocks

- **Causes** of interlocks:
  - Collusion
  - Cooptation and monitoring
  - Legitimacy
  - Career advancement
  - Social cohesion

- **Consequences** of interlocks:
  - Corporate control
  - Economic performance
  - Access to resources
Board interlocks

- **Causes** of interlocks:
  - Collusion
  - Cooptation and monitoring
  - Legitimacy
  - Career advancement
  - Social cohesion

- **Consequences** of interlocks:
  - Corporate control
  - Economic performance
  - Access to resources

Corporate networks

Figure: 400,000 largest firms globally, plotted based on latitude/longitude.
Corporate networks

Figure: Global corporate network: over 1,000,000 board interlocks.
CORPNET

- CORPNET — Corporate Network Governance: Power, Ownership and Control in Contemporary Global Capitalism

What are the features, origins and power political consequences of corporate governance networks in modern economic life?

- Nature: map and analyze the network
- Origins: uncover generating mechanisms
- Power: understand how it operates
CORPNET

- CORPNET — Corporate Network Governance: Power, Ownership and Control in Contemporary Global Capitalism
- *What are the features, origins and power political consequences of corporate governance networks in modern economic life?*
  - Nature: map and analyze the network
  - Origins: uncover generating mechanisms
  - Power: understand how it operates
- Work with Eelke Heemskerk and Javier Garcia-Bernardo
Corporate network analysis

- Apply techniques from (social) **network analysis** to corporate data
- **Nodes** represent around firms across the globe
- **Edges** denote different relationships:
  - (Undirected) **board interlocks**: shared senior level directors
  - (Directed) ownership ties based on shareholder information
- Node attributes: country, sector, performance indicators, number of employees, . . .
- Edge attributes: number of interlocks, type of shares, number of shares, ultimate share percentage, . . .
- Data source: ORBIS database
Three topics

1. Network topology & centrality
2. Community detection
3. Data quality
Network topology & centrality

Dataset

- ORBIS database (Bureau van Dijk)
- Firms listed as “large” or “very large”, and “active”
- Personal interlocks at senior management and board level
- Snapshot from December 2013
- Two-mode network of 971,891 firms and 3,272,523 top executives
- 579,924 firms did not have any interlocks
Dataset

- ORBIS database (Bureau van Dijk)
- Firms listed as “large” or “very large”, and “active”
- Personal interlocks at senior management and board level
- Snapshot from December 2013
- Two-mode network of 971,891 firms and 3,272,523 top executives
- 579,924 firms did not have any interlocks
- The remaining 391,967 nodes form the nodes in the one-mode global firm-by-firm network
## Topological properties

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td>391,967</td>
</tr>
<tr>
<td><strong>Edges</strong></td>
<td>1,711,968</td>
</tr>
<tr>
<td><strong>Density</strong></td>
<td>$2.229 \cdot 10^{-5}$</td>
</tr>
<tr>
<td><strong>Average degree</strong></td>
<td>8.746</td>
</tr>
<tr>
<td><strong>Clustering coefficient</strong></td>
<td>0.755</td>
</tr>
<tr>
<td><strong>Degree assortativity</strong></td>
<td>0.260</td>
</tr>
<tr>
<td><strong>Components</strong></td>
<td>55,616</td>
</tr>
</tbody>
</table>

**Giant component**

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td>238,859 (60.9%)</td>
</tr>
<tr>
<td><strong>Edges</strong></td>
<td>1,533,030 (89.5%)</td>
</tr>
<tr>
<td><strong>Density</strong></td>
<td>$5.374 \cdot 10^{-5}$</td>
</tr>
<tr>
<td><strong>Average degree</strong></td>
<td>12.83</td>
</tr>
<tr>
<td><strong>Clustering coefficient</strong></td>
<td>0.751</td>
</tr>
<tr>
<td><strong>Degree assortativity</strong></td>
<td>0.202</td>
</tr>
<tr>
<td><strong>Average distance</strong></td>
<td>7.775</td>
</tr>
<tr>
<td><strong>Radius</strong></td>
<td>18</td>
</tr>
<tr>
<td><strong>Diameter</strong></td>
<td>34</td>
</tr>
</tbody>
</table>
## Topological properties

### Global network

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>391,967</td>
</tr>
<tr>
<td>Edges</td>
<td>1,711,968</td>
</tr>
<tr>
<td>Density</td>
<td>$2.229 \cdot 10^{-5}$</td>
</tr>
<tr>
<td>Average degree</td>
<td>8.746</td>
</tr>
<tr>
<td>Clustering coeff</td>
<td>0.755</td>
</tr>
<tr>
<td>Degree assortativity</td>
<td>0.260</td>
</tr>
<tr>
<td>Components</td>
<td>55,616</td>
</tr>
</tbody>
</table>

### Giant component

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>238,859 nodes (60.9%)</td>
</tr>
<tr>
<td>Edges</td>
<td>1,533,030 (89.5%)</td>
</tr>
<tr>
<td>Density</td>
<td>$5.374 \cdot 10^{-5}$</td>
</tr>
<tr>
<td>Average degree</td>
<td>12.83</td>
</tr>
<tr>
<td>Clustering coeff</td>
<td>0.751</td>
</tr>
<tr>
<td>Degree assortativity</td>
<td>0.202</td>
</tr>
<tr>
<td>Average distance</td>
<td>7.775</td>
</tr>
<tr>
<td>Radius</td>
<td>18</td>
</tr>
<tr>
<td>Diameter</td>
<td>34</td>
</tr>
</tbody>
</table>
Topological distributions

Figure: Degree distribution of the giant component

Figure: Component size distribution (excluding giant component)
Topological distributions

Figure: Distance distribution

Figure: Eccentricity distribution
National networks

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GB</td>
<td>32,962</td>
<td>0.00067</td>
<td>0.356</td>
<td>0.845</td>
<td>6.63</td>
<td>0.26</td>
</tr>
<tr>
<td>US</td>
<td>24,802</td>
<td>0.00024</td>
<td>0.228</td>
<td>0.741</td>
<td>6.71</td>
<td>0.48</td>
</tr>
<tr>
<td>ES</td>
<td>11,102</td>
<td>0.00143</td>
<td>0.156</td>
<td>0.849</td>
<td>6.30</td>
<td>0.25</td>
</tr>
<tr>
<td>NO</td>
<td>8,963</td>
<td>0.00130</td>
<td>0.173</td>
<td>0.613</td>
<td>5.69</td>
<td>0.40</td>
</tr>
<tr>
<td>FR</td>
<td>8,896</td>
<td>0.00083</td>
<td>0.170</td>
<td>0.445</td>
<td>6.13</td>
<td>0.77</td>
</tr>
<tr>
<td>MY</td>
<td>7,878</td>
<td>0.00398</td>
<td>0.115</td>
<td>0.785</td>
<td>4.50</td>
<td>0.07</td>
</tr>
<tr>
<td>DE</td>
<td>7,224</td>
<td>0.00142</td>
<td>0.320</td>
<td>0.799</td>
<td>8.15</td>
<td>0.63</td>
</tr>
<tr>
<td>SE</td>
<td>6,656</td>
<td>0.00166</td>
<td>0.430</td>
<td>0.829</td>
<td>6.40</td>
<td>0.79</td>
</tr>
<tr>
<td>NL</td>
<td>6,083</td>
<td>0.00271</td>
<td>0.225</td>
<td>0.785</td>
<td>7.61</td>
<td>0.84</td>
</tr>
<tr>
<td>IN</td>
<td>5,911</td>
<td>0.00173</td>
<td>0.047</td>
<td>0.332</td>
<td>4.72</td>
<td>0.20</td>
</tr>
<tr>
<td>CA</td>
<td>5,439</td>
<td>0.00146</td>
<td>0.072</td>
<td>0.352</td>
<td>5.20</td>
<td>0.52</td>
</tr>
<tr>
<td>DK</td>
<td>4,517</td>
<td>0.00229</td>
<td>0.163</td>
<td>0.549</td>
<td>5.61</td>
<td>0.78</td>
</tr>
<tr>
<td>IT</td>
<td>4,483</td>
<td>0.00125</td>
<td>0.198</td>
<td>0.524</td>
<td>7.57</td>
<td>0.88</td>
</tr>
<tr>
<td>BE</td>
<td>3,264</td>
<td>0.00254</td>
<td>0.123</td>
<td>0.416</td>
<td>5.17</td>
<td>1.57</td>
</tr>
<tr>
<td>RU</td>
<td>2,939</td>
<td>0.00263</td>
<td>0.102</td>
<td>0.556</td>
<td>6.57</td>
<td>0.08</td>
</tr>
<tr>
<td>KR</td>
<td>2,802</td>
<td>0.00174</td>
<td>0.124</td>
<td>0.356</td>
<td>5.83</td>
<td>0.05</td>
</tr>
<tr>
<td>FI</td>
<td>2,626</td>
<td>0.00294</td>
<td>0.174</td>
<td>0.539</td>
<td>5.52</td>
<td>1.11</td>
</tr>
<tr>
<td>JP</td>
<td>2,605</td>
<td>0.00119</td>
<td>0.113</td>
<td>0.208</td>
<td>7.20</td>
<td>0.21</td>
</tr>
<tr>
<td>IE</td>
<td>2,497</td>
<td>0.001479</td>
<td>0.178</td>
<td>0.747</td>
<td>5.78</td>
<td>0.39</td>
</tr>
<tr>
<td>AT</td>
<td>2,142</td>
<td>0.00440</td>
<td>0.273</td>
<td>0.670</td>
<td>5.58</td>
<td>0.79</td>
</tr>
<tr>
<td>PT</td>
<td>2,120</td>
<td>0.00488</td>
<td>0.138</td>
<td>0.620</td>
<td>5.45</td>
<td>0.56</td>
</tr>
<tr>
<td>AU</td>
<td>1,897</td>
<td>0.00382</td>
<td>0.085</td>
<td>0.414</td>
<td>4.94</td>
<td>0.58</td>
</tr>
<tr>
<td>LU</td>
<td>1,484</td>
<td>0.00705</td>
<td>0.196</td>
<td>0.720</td>
<td>6.72</td>
<td>1.55</td>
</tr>
<tr>
<td>SG</td>
<td>1,472</td>
<td>0.00709</td>
<td>0.080</td>
<td>0.421</td>
<td>4.14</td>
<td>0.90</td>
</tr>
<tr>
<td>VN</td>
<td>1,393</td>
<td>0.00558</td>
<td>0.090</td>
<td>0.501</td>
<td>4.44</td>
<td>0.01</td>
</tr>
<tr>
<td>CH</td>
<td>999</td>
<td>0.00620</td>
<td>0.077</td>
<td>0.316</td>
<td>4.78</td>
<td>1.63</td>
</tr>
<tr>
<td>CN</td>
<td>891</td>
<td>0.00475</td>
<td>0.132</td>
<td>0.465</td>
<td>5.80</td>
<td>1.18</td>
</tr>
<tr>
<td>KY</td>
<td>642</td>
<td>0.00693</td>
<td>0.098</td>
<td>0.387</td>
<td>5.40</td>
<td>3.90</td>
</tr>
</tbody>
</table>
Findings

- Small world phenomenon
- Average node-to-node distance
  - Global network: 7.775
  - National networks: 5.692 (average) or 6.188 (weighted average)
- National footprints still visible?
Findings

- Small world phenomenon
- Average node-to-node distance
  - Global network: 7.775
  - National networks: 5.692 (average) or 6.188 (weighted average)
- National footprints still visible?
  - Competing elites
  - Globalization
Findings

- Small world phenomenon
- Average node-to-node distance
  - Global network: 7.775
  - National networks: 5.692 (average) or 6.188 (weighted average)
- National footprints still visible?
  - Competing elites
  - Globalization
- Let’s investigate more complex embeddedness measures!
Centrality

- **Node centrality**: the importance of a node with respect to the other nodes based on the structure of the network
- **Centrality measure**: computes the centrality value of all nodes in the graph
  - **Degree centrality**: number of connections
  - **Closeness centrality**: average distance to all other nodes
  - **Betweenness centrality**: relative number of times a node is on a shortest path
- But what is the ground truth to verify these measures?
Centrality

- **Node centrality**: the importance of a node with respect to the other nodes based on the structure of the network
- **Centrality measure**: computes the centrality value of all nodes in the graph
  - **Degree centrality**: number of connections
  - **Closeness centrality**: average distance to all other nodes
  - **Betweenness centrality**: relative number of times a node is on a shortest path
- But what is the ground truth to verify these measures?
  - Hard to say!
**Centrality**

- **Node centrality**: the importance of a node with respect to the other nodes based on the structure of the network
- **Centrality measure**: computes the centrality value of all nodes in the graph
  - **Degree centrality**: number of connections
  - **Closeness centrality**: average distance to all other nodes
  - **Betweenness centrality**: relative number of times a node is on a shortest path
- But what is the ground truth to verify these measures?
  - Hard to say!
  - Correlate with **firm prominence** (revenue)?
Centrality measures compared

**Figure:** Degree, closeness and betweenness centrality

Source: “Centrality” by Claudio Rocchini, Wikipedia File:Centrality.svg
Global vs. National centrality

Global
1. US AT&T INC.
2. US 7-ELEVEN INC.
3. GB ROYAL DUTCH SHELL
4. GB ERNST & YOUNG EUROPE
5. KR SAMSUNG ELECTRONICS
6. GB PRICEWATERHOUSECOOPERS
7. CH RAiffeisen SCHWEIZ
8. GB KPMG EUROPE
Global vs. National centrality

Global
1. US AT&T INC.
2. US 7-ELEVEN INC.
3. GB ROYAL DUTCH SHELL
4. GB ERNST & YOUNG EUROPE
5. KR SAMSUNG ELECTRONICS
6. GB PRICEWATERHOUSECOOPERS
7. CH RAIFFEISEN SCHWEIZ
8. GB KPMG EUROPE

Great Britain
1. GB ERNST & YOUNG EUROPE
2. GB PRICEWATERHOUSECOOPERS
3. GB KPMG EUROPE
4. GB ROYAL DUTCH SHELL
5. GB DELOITTE
6. GB JP MORGAN
7. GB EASYJET
8. GB DLA PIPER INTERNATIONAL
Global centrality

Table: Correlation between centrality measures and with firm prominence (revenue), $n = 238,859$.

<table>
<thead>
<tr>
<th></th>
<th>Betweenness</th>
<th>Closeness</th>
<th>Degree</th>
<th>Eigenvector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Betweenness</td>
<td>1.000</td>
<td>0.430</td>
<td>0.521</td>
<td>0.356</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.430</td>
<td>1.000</td>
<td>0.495</td>
<td>0.902</td>
</tr>
<tr>
<td>Degree</td>
<td>0.521</td>
<td>0.495</td>
<td>1.000</td>
<td>0.498</td>
</tr>
<tr>
<td>Eigenvector</td>
<td>0.356</td>
<td>0.902</td>
<td>0.498</td>
<td>1.000</td>
</tr>
<tr>
<td>Firm prominence</td>
<td>0.192</td>
<td>0.109</td>
<td>-0.046</td>
<td>0.064</td>
</tr>
</tbody>
</table>
**National centrality**

![Bar chart showing the correlation between firm prominence (revenue) and national centrality.](image)

**Figure:** Correlation between firm prominence (revenue) and national centrality

<table>
<thead>
<tr>
<th>Country</th>
<th>Betweenness centrality</th>
<th>Closeness centrality</th>
<th>Degree centrality</th>
<th>Eigenvector centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AU</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ES</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KY</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LU</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MX</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MY</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RU</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
National vs. global centrality

- **Centrality persistence**: correlation between global centrality (in the full network) and national centrality (in a partition)
Centrality persistence

Figure: Centrality persistence for the 35 largest countries.
Centrality persistence

Figure: Betweenness centrality persistence vs. normalized log(GDP).
National vs. global centrality

- **Partition ranking dominance**: the ranking of a partition within the full network.

  Assume \( S \subseteq V \) in a graph \( G = (V, E) \).

  Assume that a node \( v \in S \) according to some centrality ranking has rank \( r(v) \in [0, |V|] \) in the full ranking of all nodes in \( V \).

  Partition ranking dominance \( pcr(S, V) \) is then defined as:

  \[
  pcr(S, V) = 0.5 - \frac{\sum_{v \in S} r(v)}{|S| \cdot |V|}
  \]

  - value > 0 means the partition is less central than expected
  - value < 0 means it is more central than expected
Partition ranking dominance

Figure: Partition ranking dominance (based on betweenness centrality).
Community detection

Community detection

- **Community**: set of nodes connected more strongly with each other than with the rest of the network
- Community detection algorithms:
  - Clique-based methods
  - Hierarchical clustering
  - Divisive algorithms (centrality-based)
  - **Modularity maximization** algorithms
- Country network: aggregate firms from the same country
Community detection

Figure: Communities: node subsets connected more strongly with each other
Modularity

- **Modularity**: numerical value indicating the quality of a division of a network into communities
- **Community**: subset of nodes for which the fraction of links inside the community is higher than expected in a random network
- Modularity $Q \in [0, 1]$
- Resolution parameter $r$ indicating how “tough” the algorithm should look for communities
- Algorithms optimize the modularity score $Q$ given some $r$ (using hill climbing, heuristics, genetic algorithms and many more optimization techniques)

Community detection
Community detection
Community detection
Community detection
Community detection
Community detection
Community detection
Community detection

Figure: Communities: node subsets connected more strongly with each other

- Communities in corporate networks have a **regional** character and **financial ties** are clearly visible
- Historical events and cultural similarities between countries correlate with interlocks
Community detection

Figure: Communities: node subsets connected more strongly with each other

- Communities in corporate networks have a **regional** character and **financial ties** are clearly visible
- Historical events and cultural similarities between countries correlate with interlocks
- Outliers and effects of randomization
Computing infrastructure

- Server grade hardware
- Dual high-frequency CPU architecture with $2 \times$ Intel Xeon (Haswell) E5-2643, 6 cores, 12 threads, 3.4GHz
- Memory: DDR4-2133 RAM, $24 \times 64GB = 1536GB = 1.5\text{TB}$
- Storage: 7TB solid state disk (SDD) storage in RAID6
- 1GBit uplink to the world

Made possible by the High Performance Computing and Networking (HPCN) fund (summer 2015 call) of the University of Amsterdam.
Data quality
Data quality

- Previous dataset was from September 2013
- CORPNET: study all firms
- More than 200 million firms
- Are all firms equally important?
- Do we have all the firms?
- What is the quality of the data?
Data quality

- Data quality
  - Accuracy: the data is true
  - Consistency: data remains clear and verifiable over time
  - Integrity: data has not suffered from corruption
  - Completeness: do we have all the data?

- We “found” that the Spanish market size was ten times larger than the US market: one outlier in the data.
Average operating revenue

Figure: Observed average revenue per country for 200 million firms
Data quality

- Assess firm data quality based on comparing intrinsic factors of countries using:
  - Worldbank data on GDP per capita for each country
  - Eurostat data on the number of firms in each county
  - Distribution of sum of revenues per country in our data
Data quality

Figure: Richer countries have larger firms
Data quality

**Figure:** Richer countries have better quality
Data quality

- Rich countries have higher average revenue, but better quality, which decreases the observed average (hard to decouple).
- We are interested in the real average (given complete data):
  1. Real average $\propto \frac{\text{GDP}}{\text{number of firms}}$
  2. Calculate the effect of intrinsic factors and extrapolate to other countries
  3. Calculate the quality of our global firm data
Data quality

- The distribution of firm operating revenues follows a lognormal distribution for 95% of firms, with consistent variance.
- Larger firms are well-represented. Richer countries have higher data quality. Higher quality decreases the observed average.

**Figure:** Lognormal distribution and addition of firms
Data quality

Figure: Observed average revenue

Figure: Estimated average revenue
Figure: log(predicted observed) = 3.15 log(estimated real) + log(completeness) − 1.05

Figure: Actual completeness of our data
Completeness per country
Other directions

- Revolving doors
- Top income compensation
- Firms in occupied territories
- Public tender, procurement
- Relation with patent networks
- Exchange Traded Funds (ETFs)
- Analysis of particular national networks
Conclusion

- Big corporate network data provides interesting insight in firm power and control across the globe.
- Topological properties, centrality analysis and community detection reveal regional patterns in the global network.
- Interpretation of measures is crucial and depends on data quality.
- We understand the completeness of our 200 million firm dataset, now we can assess the effect on the network.
Conclusion

- Big corporate network data provides interesting insight in firm power and control across the globe
- Topological properties, centrality analysis and community detection reveal regional patterns in the global network
- Interpretation of measures is crucial and depends on data quality
- We understand the completeness of our 200 million firm dataset, now we can assess the effect on the network
- **CORPNET** has a challenging yet exciting time ahead!
  Website: [http://corpnet.uva.nl](http://corpnet.uva.nl)
- We are open to sharing data, best practices and ideas!
Thank you!

Questions?

http://franktakes.nl
http://corpnet.uva.nl