Leveraging Big Data to Help Competition Agencies Tackle Anticompetitive Behavior

Georgiana Pop and Andrew Whitby

SUMMARY
Anticompetitive practices have been found to yield negative effects on productivity growth in developed and developing economies. Cartels across the world, for instance, negatively impact consumer welfare through price overcharges in the order of billions of dollars. But detecting these practices requires a better understanding of the nature of such anticompetitive practices by market operators across various markets and jurisdictions. This is not an easy task. Currently, competition authorities and researchers who study effects of anticompetitive behavior may have to gather this data manually, which is both time-consuming and inefficient.

In response, this project aimed to develop a database of key decisions by the competition authorities relating to anti-competitive practices. Using machine learning and web scraping techniques, the project automates the collection and organization of data from sixteen pilot countries. The database would serve as an essential infrastructure for future visualization and analyses to identify signals of anti-competitive behavior.
**CHALLENGE**

Data on anticompetitive practices is an integral part of Competition Authorities’ work. For example, in the case of cartels, such data may cover data on market characteristics, specific regulations, conduct, decisions, and sanctions imposed. But many times competition authorities do not currently have ready access to cross-market, cross-jurisdictional data on such anti-competitive practices. This is because this type of data is often collected manually from public sources such as competition authorities’ websites, media, and specialized international organizations and consultancies. Moreover, building and maintaining a repository of this data “by hand” is time consuming – and can yield imperfect results.

**INNOVATION**

Using machine learning techniques and web scraping, this project aimed to automate the collection and organization of key information on anti-competitive practices from sixteen pilot countries.

The database would allow competition authorities to better understand, detect, and take actions against actors that systematically engage in anti-competitive practices. At the World Bank Group, it would allow the Competition Policy Team to further develop its analytical tools and systematize its corpus of evidence on the effects of anticompetitive practices, including cartels in developing economies.

Further, this initiative could potentially be expanded to the analysis of various anticompetitive practices as well the implementation of other policies, e.g., investment policy and state-aid control policy.
RESULTS
The team initially started out with a set of countries where competition authorities had published documents – in either English or Spanish – relating to anti-competitive practices. This initial set was narrowed down to sixteen pilot countries where a comprehensive set of records on past decisions was available. The pilot focused on the following sixteen countries: Albania, Botswana, India, Moldova, Romania, Uruguay, Argentina, Chile, Macedonia, Pakistan, Serbia, Bosnia, Costa Rica, Malaysia, Peru, and Seychelles.

In the first phase of the project, the team worked with a partner organization with technical expertise in machine learning and web-scraping to scan thousands of pages of documents and gather semi-structured data on anticompetitive practices. This involved creation of an algorithm that picked out relevant information from 5000 documents – including the actors involved, type of anti-competitive practice, products/markets affected, type of anticompetitive practices, decision taken, and sanctions, if any. This phase required the imposition of both automated and manual checks and troubleshooting to ensure credibility and quality of data collected.

The second stage, which is in progress, involves the extraction and structuring of this information so that it is ready for analysis.

In the third phase, the team hopes use this database to gather insights on anti-competitive behavior. This could potentially include: (1) snapshots of anti-competitive behavior by product/market, (2) network mapping and analysis to identify companies linked to previous record-holders across different markets, jurisdictions, and geographies, and (3) likelihood analysis of potential anticompetitive behavior for companies with links or previous records, etc.
LESSONS LEARNED

1. **Allocate time and resources to communicate domain knowledge with technical experts:** Machine learning and web scraping expertise needs to be complemented by transfer of domain-specific knowledge to technical experts. For this, the team found it important to allocate enough time and resources to make sector-specific terms and analytical approaches/practices explicit to technical experts.

2. **Human oversight is crucial in ensuring quality:** There is no perfect analytical technique when it comes to unstructured data, because so much depends on the quality of documents available online. Since there is a lot that needs to be taught to the machine, the human factor is very crucial in ensuring quality of outputs.

   For example, while a human expert may intuitively place the terms “cartel”, “price agreement”, and “horizontal agreement” in the same category of anti-competitive practice, a machine has to be explicitly taught that this is the case. Breaking down what a human does into concrete steps and feeding it to the machine takes time—both in terms of algorithm design and in terms of cross-checking, so the process cannot be entirely automated at this time.

3. **Anticipate adjustments to time-frame:** From database creation to completing complex analyses, each phase of the project builds on the previous one. The team found it important to factor in additional time so that quality can be assured adequately throughout the project lifecycle.