Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Maestría en Exploración de Datos y Descubrimiento del Conocimiento

Tesis a ser presentada para obtener al título de Magíster en Explotación de Datos y Descubrimiento de Conocimiento

Plan de Tesis de Maestría en Minería de Datos y Descubrimiento de conocimiento

Autor: Esp. Diego Kozlowski

Directora de Tesis:
Dra. Viktoriya Semeshenko
29 de noviembre de 2018

Índice general

1. Resumen

El presente plan de trabajo propone como proyecto de tesis el desarrollo de un conjunto de herramientas basadas en teoría de grafos y modelos generativo bayesianos, aplicadas a la caracterización del comercio internacional. En particular, se propondrán diferentes construcciones de grafos para analizar distintos aspectos de las relaciones comerciales internacionales, tales como la importancia de los países y la complejidad de sus matrices productivas. También se propone la implementación de modelos originalmente inspirados en el Text Mining, para el agrupamiento automático de los productos y la caracterización del rol de los países en la división internacional del trabajo.

2. Tema de investigación

El análisis del comercio internacional constituye una de las áreas de estudio más importantes de la investigación económica. Desde los comienzos de la economía política clásica constituye un tema de preocupación por sus fuertes implicancias (Ricardo, 2007). Por su parte, el registro de la información referente al comercio entre países también se remonta en el tiempo.

El comercio internacional se puede pensar como la interacción de cientos de países (N), miles de productos (P), a lo largo de cientos de años (Y), en diversas direcciones (exportaciones e importaciones). Es decir, el problema existe en un espacio de dimensión $\mathcal{R}^{N*P*Y*T*2}$. Esta estructura compleja de interacciones se suele caracterizar con grandes agregados sobre alguna dimensión, perdiendo la riqueza de la información sobre las demás. El análisis tradicional de la información generada carece de las herramientas necesarias para hacer frente a los grandes volúmenes de datos generados por el comercio internacional en la actualidad. Históricamente, los indicadores sintéticos del área se resumen en volumen y masa de dinero comerciada, partiendo del total mundial, hacia desagregaciones por región, país y sector económico en cuestión (WTO, 2017). Una parte importante del análisis que se propone superar la generalidad del total comerciado entre dos países, lo hace al concentrarse en un sector particular de la economía, intentando reconstruir las cadenas globales de valor para dicha rama (Giuliani, Pietrobelli y Rabellotti, 2005). Por su parte, también se han propuesto modelos que explican el flujo bilateral entre países basados en el tamaño de los países en cuestión, y las distancias que los separan (Head y Mayer, 2014).

A su vez, el análisis del comercio internacional se interesa no sólo por los flujos comerciales en sí, sino también por lo que ellos reflejan respecto de la División Internacional del Trabajo. Esto es, el lugar que ocupan los diferentes países en el mercado mundial, tanto respecto a su importancia relativa, como también su especialización en una cierta canasta de productos (Fröbel, Heinrichs y Kreye, 1978).

* * *

A la vez que aumenta el volumen de información persistida respecto del comercio internacional, también se facilita el acceso a técnicas de análisis de datos que requieren un mayor poder de cómputo, y que por lo tanto eran impensadas como herramientas de estudio en épocas anteriores. En este sentido, surge la posibilidad de complementar el análisis tradicional del comercio mundial con nuevas herramientas que guardan la posibilidad de echar nueva luz sobre fenómenos largamente estudiados.

El presente trabajo se propone utilizar técnicas propias del análisis de grafos para caracterizar el comercio internacional. Su modelización como una red compleja permite la construcción de medidas de resumen que, sin abandonar una mirada holística de la problemática, logren dar cuenta de una mayor

complejidad que las métricas tradicionales de dicha área temática. Para ello, se evaluarán alternativamente modelos de comercio agregado entre países (Fagiolo, Reyes y Schiavo, 2007), así como también grafos bipartitos para analizar el comercio a nivel producto (Hidalgo y col., 2007), a la vez que se observará la evolución temporal como una evolución en las medidas de centralidad de cada nodo en cada punto del tiempo. Por último, también se propone como técnica alternativa de análisis una aplicación de Latent Dirichlett Allocation Models (David M. Blei Andrew Y. Ng, 2003) a la esfera del comercio internacional.

Los modelos basados en grafos resultan particularmente interesantes para dilucidar el lugar que los países ocupan en la División Internacional del Trabajo, ya que el concepto de centralidad de la teoría de redes aplica directamente a este campo de estudio. Sin embargo, la especialización productiva de los países, es decir qué subconjunto de productos exportan e importan con mayor frecuencia, representa una complejidad mayor. Es es debido a la multiplicidad e interrelación del espacio de producción. Los nomencladores tradicionales consideran miles de productos diferentes, los cuales comparten mayores o menores similitudes entre sí, en varias dimensiones distintas. Dos productos pueden ser considerados similares según la rama de la producción a la que pertenecen, las materias primas de las que están hechos, el nivel de complejidad que poseen o el valor agregado que generan (Molinari y Angelis, 2016). La caracterización del lugar que un país ocupa en el mercado mundial no puede simplemente reducirse a los volúmenes comerciados, sino que debe considerar qué productos comercia, y las interrelaciones mencionadas.

Es por ello que para el presente trabajo se propone en primer lugar un grafo bipartito de países y productos, cuya proyección como un grafo de producto permitiría caracterizar las interacciones entre éstos. De forma complementaria, se propondrá una técnica experimental, basada en el modelo gráfico de Latent Dirichlett Allocation, con el cual se buscara construir una variedad del espacio del producción sobre la que yacen dimensiones latentes que definen la relación entre los productos, a la vez que se reconstruye la canasta exportadora de los países respecto a dicho espacio latente.

3. Estado del arte

No son pocos los autores que utilizan el análisis de grafos para desarrollar intuiciones respecto al comportamiento económico. En la literatura económica general, Jackson (2008) presenta múltiples implementaciones de la teoría de grafos para el análisis económico. En la literatura orientada al comercio internacional, existen sendos intentos de representar estos flujos como una red compleja. Ya sea desde la teoría de la información (Bhattacharya y col., 2008), como una herramienta de modelización de los fenómenos económicos (Fan y col., 2014), para analizar las relaciones de centro-periferia (Fagiolo, Reyes y Schiavo, 2010), o bien para realizar una descripción del estado del comercio internacional en un momento dado (Chow, 2013). Por su parte, cabe destacar los trabajos de Hidalgo (Hidalgo y col., 2007; Hidalgo, 2009; Hidalgo y Hausmann, 2009), donde se elabora un grafo bipartito a partir del comercio bilateral a nivel producto, para desarrollar el concepto de complejidad económica de los productos y las naciones. A su vez, como resultado de lo anterior, se obtiene un mapa de la complejidad de los productos. Este *espacio de productos* surge de la proyección de un grafo bipartito entre países y productos, en el cual se define una arista entre un país y un producto si el país en cuestión exhibe *Relative Comparative Advantages*, una medida que denota la propensión de ese país a componer su canasta exportadora por dicho producto. Dada la amplitud de los temas a abordar, tanto desde la perspectiva económica, como desde la teoría de grafos, éste se presenta como un campo abierto para la investigación, con sendas aristas aún por recorrer.

En paralelo a lo anterior, el análisis estadístico de la composición de las canastas exportadoras de los países se realiza tradicionalmente agregando los diferentes productos sobre la base de nomencladores realizados especialmente por comités de expertos de organismos internacionales (ONU, 2006; OECD, 2017). En dichos nomencladores se construye una estructura jerárquica, donde cada dígito se corresponde con alguna dimensión en la cual se puede trazar una distancia entre los diferentes productos, por ejemplo por rama de actividad, complejidad técnica, etc. A su vez, sobre la base de las clasificaciones estandarizadas, muchos autores construyen sus propias clasificaciones buscando alguna dimensión específica. Por ejemplo Lall (2000) realiza una clasificación de productos según si su tecnología es baja, media-baja, media-alta y alta. Por su parte Molinari y Angelis (2016) realizan una clasificación que considera las cadenas de valor, es decir aquellos eslabonamientos productivos que van desde las materias primas hasta el producto final. A su vez, Flôres Jr (2008) propone una clasificación de los productos según su uso como bienes de capital, productos primarios, etc. Todas estas construcciones se realizan de forma manual, y se basan en el conocimiento de expertos en la temática. Es por esto último, que resulta de interés la elaboración de

técnicas de clasificación de productos que sean automáticas, y en lugar de basarse en el conocimiento de especialistas, se fundamenten en la información disponible.

En paralelo, existen múltiples técnicas desarrolladas en el campo del Text Mining que buscan resumir automáticamente los textos. En particular, la técnica conocida como Lantent Dirichlet Allocation Models, o Topic Modelling, desarrollada por David M. Blei Andrew Y. Ng (2003). El principal objetivo de estos modelos es encontrar los tópicos de los que habla un texto, considerado como parte de un Corpus, con una cantidad definida de tópicos desconocidos. Para una intuición del modelo sobre el que se construye dicho objetivo, es necesario considerar, en el texto, cada palabra como una dimensión, y cada documento como un punto en un espacio de muy alta dimensionalidad, definido por el tamaño del vocabulario. Topic Modeling considera que sobre este espacio de alta dimensionalidad yace una variedad, de dimensionalidad mucho menor. Ésta estaría constituida por los ejes temáticos de los documentos. dichas dimensiones latentes se construyen como una distribución sobre el espacio original de palabras. Modelos con estas características permiten a la vez obtener información sobre las temáticas subyacentes de un Corpus, así como también conocer los principales ejes temáticos de cada texto.

El presente trabajo se propone tomar por base el desarrollo realizado por David M. Blei Andrew Y. Ng (2003) y reconfigurar el modelo de forma tal que brinde una nueva mirada sobre el comercio internacional. Para ello, se recupera el modelo gráfico propuesto por los autores, dejando de lado las intuiciones respecto al procesamiento de texto. Si bien el trabajo original de los autores explicita que el modelo propuesto no necesariamente debe atenerse a los margenes temáticos del análisis de textos, no abundan trabajos que recuperen dicho modelo para otros tipos de datos. La estructura de variables aleatorias con relaciones jerárquicas propuesta por el modelo original posee una analogía ideal para el presente trabajo. Mientras en el modelo original se plantea un proceso generativo con la jerarquía *Corpus-Documento-dimensión latente-palabra*, en el presente trabajo se propone la estructura *Mundo-País-dimensión latente-exportaciones*. Esta analogía nos permite hacer uso de las técnicas inferenciales propuestas en el modelo original (Hoffman y col., 2013) y obtener estimaciones del espacio latente, tanto respecto de su configuración como de su distribución en los otros niveles del modelo. Es decir, así como en el modelo original se parte de las palabras de los documentos para caracterizar los tópicos de los que habla y la distribución de tópicos en cada documento, en la presente propuesta se busca partir de las exportaciones de cada país para obtener la distribución de la dimensión latente del comercio internacional, las ramas o componentes, así como también obtiener información respecto de en qué componente se especializa cada país en su comercio respecto del mundo. De esta forma se obtendría una respuesta novedosa para la pregunta clásica en el estudio del comercio internacional sobre la especialización productiva de los países, arriba mencionada.

4. Objetivos

Los aportes que se espera realizar se pueden resumir en los siguientes puntos:

Objetivo general: Reconocer, a partir de la modelización del comercio mundial en una red compleja, los patrones generales de la economía mundial.

Objetivos específicos:

1. Desarrollar una metodología de análisis del comercio internacional basado en grafos.
2. Analizar comercio bilateral total entre países para el período comprendido entre 1996 y 2016 y reconocer los patrones generales del comercio mundial.

3. Sobre la base del modelo anterior, analizar la evolución de la economía mundial en el período 1950-2000, en busca de evidencias de la denominada Nueva División Internacional del Trabajo.

4. Analizar el comercio bilateral a nivel producto de los países de Sudamérica, en busca de posibles patrones de integración regional.

5. Elaborar un modelo de *Latent Dirichlet Allocation* para el comercio de productos por países

5. Antecedentes personales en el tema

Entre los antecedentes personales en la temática cabe destacar:

- **Tesis de especialización en Data Mining** (Kozlowski, 2018a): En la misma se elaboró una primera versión del objetivo específico 1. Debido a falta de acceso a la información, la misma fue recuperada de la página web de la Organización Mundial del comercio (WTO) mediante métodos de scrapping, dónde existe la posibilidad de que haya una porción de datos no recuperada. En la actualidad, se cuenta con acceso oficial a dicha base de datos, por lo que se puede enriquecer el análisis.

- **Presentación en Young Scholar Initiative 2018** (Kozlowski, 2018b): Se avanzó en los objetivos específicos 2 y 3. A su vez, en dicha conferencia se recibieron valiosas críticas respecto a la metodología propuesta, que permitirían robustecer la misma.

- **Presentación en LatinR 2018** (Kozlowski, 2018c).

- **Presentación en el seminario ECON 2018** : En octubre 2018 se presentó un seminario sobre los avances de una investigación que se lleva a cabo desde mayo 2018 junto a las profesoras Andrea Molinari y Viktoriya Semeshenko del Instituto de Investigación en Economía Política (IIEP)-FCE-UBA, para elaborar una representación del comercio bilateral en América Latina, a partir de un grafo bipartito. En las colaboraciones con dicho grupo se avanza con los objetivos específicos 4 y 5.

6. Plan de Trabajo

Etapas de desarrollo de la tesis:

1. **Revisión bibliográfica.** Investigación del estado del arte. En particular:
   - La utilización de la teoría de grafos en las ciencias sociales.
   - Las representaciones del comercio internacional como una red compleja
   - Las propuestas de Hidalgo (Hidalgo y col., 2007; Hidalgo y Hausmann, 2009; Hidalgo, 2009) para el análisis de la complejidad del espacio de productos.

2. **Recopilación de la información.** Descarga de datos de la página de la Organización del Comercio internacional, y búsqueda de fuentes alternativas, de mayor extensión temporal.

3. **Elaboración de un grafo del comercio bilateral agregado para el año 2016** Análisis de las decisiones metodológicas que involucran la conceptualización de la red, y los primeros resultados obtenidos. En base a la información provista por la *WTO*

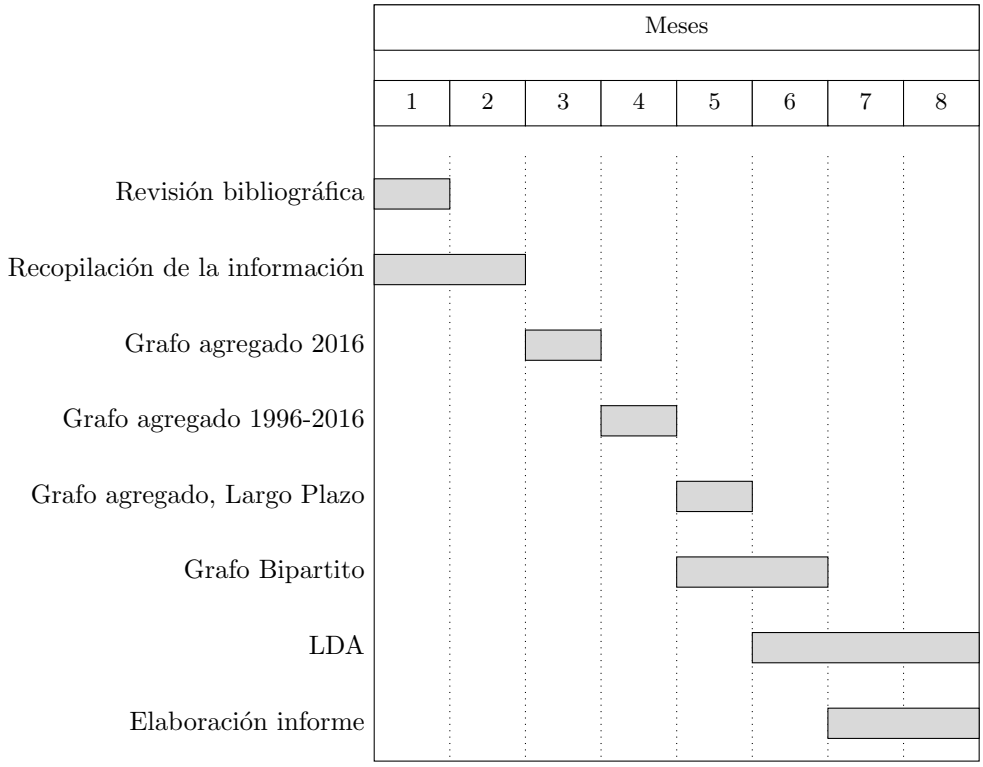
4. **Elaboración grafo de comercio bilateral agregado para el período 1996-2016.** Análisis de la evolución de las métricas que caracterizan al grafo y a los nodos en el período reciente. En base a la información provista por la *WTO*

5. **Análisis de los movimientos de largo plazo en la red.** Búsqueda de elementos confirmatorios o que refuten la tesis de la nueva división internacional del trabajo. En base a la información provista por Gleditsch (2002)

6. **Elaboración del grafo bipartito a nivel producto.** Análisis de cercanía del espacio de productos.

7. **Desarrollo de LDA para comercio internacional.** propuesta metodológica y diseño de implementación. Análisis de las dimensiones latentes desarrolladas por el modelo, y la evolución en el tiempo de la participación de dichas dimensiones en la canasta exportadora de los países.

8. **Análisis de resultados obtenidos y preparación del informe final.**



Referencias

[1] K. Bhattacharya y col. "The international trade network: Weighted network analysis and modelling". En: *Journal of Statistical Mechanics: Theory and Experiment* 2008.2 (2008). ISSN: 17425468. DOI: 10.1088/1742-5468/2008/02/P02002. arXiv: 0707.4343.

[2] William Chow. "An Anatomy of the World Trade Network". En: (2013), págs. 1-20.

[3] Michael I Jordan David M. Blei Andrew Y. Ng. "Latent Dirichlet Allocation". En: *Journal of Machine Learning Research* 3 (2003), págs. 993-1022.

[4] Giorgio Fagiolo, Javier Reyes y Stefano Schiavo. "The evolution of the world trade web: A weighted-network analysis". En: *Journal of Evolutionary Economics* 20.4 (2010), págs. 479-514. ISSN: 09369937. DOI: 10.1007/s00191-009-0160-x.

[5] Giorgio Fagiolo, Javier Reyes y Stefano Schiavo. "The Evolution of the World Trade Web Giorgio". 2007.

[6] Ying Fan y col. "The state's role and position in international trade: A complex network perspective". En: *Economic Modelling* 39 (2014), págs. 71-81. ISSN: 02649993. DOI: 10.1016/j.econmod.2014.02.027. URL: http://dx.doi.org/10.1016/j.econmod.2014.02.027.

[7] Renato G Flôres Jr. "The World Fragmentation of Production and Trade Concepts and Basic Issues". En: *International WorkshopÏntegração Produtiva-Lições da Ásia e Europa para o MERCOSUL", CEPAL/Brasil*. 2008, pág. 7.

[8] Folker Fröbel, Jürgen Heinrichs y Otto Kreye. "The new international division of labour". En: *Information (International Social Science Council)* 17.1 (1978), págs. 123-142.

[9] Elisa Giuliani, Carlo Pietrobelli y Roberta Rabellotti. "Upgrading in global value chains: lessons from Latin American clusters". En: *World development* 33.4 (2005), págs. 549-573.

[10] Kristian S Gleditsch. "Expanded Trade and GDP Data". En: *Journal of Conflict Resolution* 46 (2002), págs. 712-24.

[11] Keith Head y Thierry Mayer. "Gravity Equations: Workhorse,Toolkit, and Cookbook". En: *Handbook of International Economics* 4 (2014), págs. 131-195. ISSN: 1573-4404. DOI: 10.1016/B978-0-444-54314-1.00003-3. URL: https://www.sciencedirect.com/science/article/pii/B9780444543141000033?via{\%}3Dihub.

[12] C. A. Hidalgo y col. "The product space conditions the development of nations". En: *Science* 317.5837 (2007), págs. 482-487. ISSN: 00368075. DOI: 10.1126/science.1144581. arXiv: 0708.2090.

[13] César Hidalgo y Ricardo Hausmann. "The building blocks of economic complexity". En: *Proceedings of the National Academy of the Sciences of the United States of America* 106.26 (2009), págs. 10570-10575. ISSN: 0027-8424. DOI: 10.1073/pnas.0900943106. arXiv: 0909.3890.

[14] César A. Hidalgo. "The Dynamics of Economic Complexity and the Product Space over a 42 year period". En: *CID Working Paper* 189.189 (2009), pág. 20. ISSN: 6507247197.

[15] Matthew D Hoffman y col. "Stochastic variational inference". En: *The Journal of Machine Learning Research* 14.1 (2013), págs. 1303-1347.

[16] Matthew O Jackson. "Social and Economic Networks". En: *Network* March (2008), págs. 14-16. ISSN: 0691148201. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

[17] Diego Kozlowski. *Descripción del comercio internacional utilizando un modelo de redes complejas*. 2018.

[18] Diego Kozlowski. "Descripción del comercio internacional utilizando un modelo de redes complejas". En: *YSI Latinamerica*. 2018.

[19] Diego Kozlowski. "Descripción del comercio internacional utilizando un modelo de redes complejas". En: *LatinR*. 2018.

[20] Sanjaya Lall. "The Technological structure and performance of developing country manufactured exports, 1985-98". En: *Oxford development studies* 28.3 (2000), págs. 337-369.

[21] Andrea Molinari y Jesica Yamila de Angelis. "Especialización y complementación productiva en el MERCOSUR: un Enfoque de Cadenas Productivas de Valor". En: (2016).

[22] OECD. "Harmonised System 2012". En: (2017). DOI: https://doi.org/https://doi.org/10.1787/6a608f64-en. URL: https://www.oecd-ilibrary.org/content/data/6a608f64-en.

[23] ONU. "Standard International Trade Classification Revision 4". En: *Statistical Papers Series M No* 34 (2006).

[24] David Ricardo. *Principios de economía política y tributación*. Ed. por Claridad. Buenos Aires, 2007.

[25] WTO. *World Trade Statistics Review 2017*. Inf. téc. 2017, pág. 181.