The Web of Innovation:
Using Website Data to Understand How Firms Innovate

Presenter: Sarah Kelley
Collaborators: Evgeny Klochikhin, PhD | Sanjay K. Arora, PhD | Sarvothaman Madhavan
Global Tech Mining Conference
Atlanta, Georgia
October 9, 2017
Previous work and motivations

- Website data on firms is freely available; prior research has found that many innovative firms have websites but lack patents (Yin et al., 2016)
- Moreover, survey response rates for firms continue to fall (Baruch, 1999)
- Working with websites presents specific challenges to social scientists who must be increasingly adept at processing unstructured data and operationalizing valid and reliable variables (Arora et al., 2016)
- Previous work has attempted to scrape firm websites *and*:
  - Validate the operationalization of variables to assess internal validity concerns (Gok et al., 2015)
  - Cluster firms by type of firm using simple keyword based approaches (Arora et al., 2013)
  - Measure firm change over time (as a proxy for firm “seizing”) as an endogenous predictor of performance (Arora et al., 2017)
Our contributions

• Our work seeks to improve applications of using website data for studying innovation

• This presentation focuses on our method for building a sample of innovative (inventive) firms whose websites can be mined and analyzed

• In particular, we explore narrative construction and detection on firm websites within a comparative framework setting
Data sources and sample frame definition

1. Firms that invent
   - Search for assignees in patents
     - Utility patents in three sectors: nanotechnology, synthetic biology, and renewable energy

2. ...and that are small
   - Get URLs and other firm characteristics from SAM.gov
     - Check for firm size using sam.gov and obtain URLs

3. ...and that have websites
   - Capture firm websites
     - Collect visible text from firm websites
Patents querying approach

• Using prior published work search terms are obtained for
  • nanotechnology sector (Arora et. al, 2012)
  • green technology sector (Shapiro, Klochikhin et. al, 2013)

• For synthetic biology sector, terms are obtained from wikipedia by using below steps:
  • A base list of terms is obtained from prior published work (Raimbault et. al, 2016)
  • From the Wikipedia page (if present) of each of the terms, all outgoing links are gathered
  • The above list is reviewed to retain terms that are deemed relevant by the researchers
  • Link extraction and review is repeated on retained terms to obtain researchers' terms list
  • The researchers’ final terms list is reviewed by a domain expert to correct for false positives and false negatives

• Patents, and consequently assignee firms, are selected by searching for the final list of terms in the patent database (in title and abstract)

• Using data provided United States System for Award Management (SAM), firms are filtered based on their small business status and their corporate URLs are obtained
Who are these firms?

<table>
<thead>
<tr>
<th>Stat</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of organization in SAM list</td>
<td>620,206</td>
</tr>
<tr>
<td>Small businesses in SAM list</td>
<td>347,249</td>
</tr>
<tr>
<td>Total number of patents</td>
<td>6,200,505</td>
</tr>
<tr>
<td>Utility patent containing the terms</td>
<td>Green Sector: 2,436</td>
</tr>
<tr>
<td>Patents with US assignee information</td>
<td>Green Sector: 1,576</td>
</tr>
<tr>
<td>Unique number of assignee organization</td>
<td>Green Sector: 607</td>
</tr>
<tr>
<td>Patent assignees org in SAM small business list</td>
<td>Green Sector: 41</td>
</tr>
</tbody>
</table>
## Assignee Representation in SAM by Patent Category

<table>
<thead>
<tr>
<th></th>
<th>Over Represented</th>
<th>Under Represented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green Technology</td>
<td>Basic electric elements</td>
<td></td>
</tr>
<tr>
<td>Synthetic Biology</td>
<td></td>
<td>Physical or chemical processes or apparatus in general; Climate change</td>
</tr>
<tr>
<td>Nanotechnology</td>
<td>Medical or veterinary science; hygiene</td>
<td></td>
</tr>
</tbody>
</table>
## Comparison of Small and Large Businesses in SAM

<table>
<thead>
<tr>
<th></th>
<th>More Small Businesses</th>
<th>More Large Businesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green Technology</td>
<td>Lighting; Organic/Inorganic chemistry; Beers, spirits &amp; wine</td>
<td>Electric elements; Climate change</td>
</tr>
<tr>
<td>Synthetic Biology</td>
<td>Medical or veterinary science; hygiene; Organic chemistry; biochemistry</td>
<td></td>
</tr>
<tr>
<td>Nanotechnology</td>
<td>Biochemistry working of plastics Cements; concrete; artificial stone; ceramics; organic compounds and their chemical preparation</td>
<td></td>
</tr>
</tbody>
</table>
Webcrawling details

• 195 firm websites across three sectors (178 unique)
• Clean urls and extract visible text data using Python/Beautiful Soup
• 162 website homepages successfully parsed
  • 23 in green goods, 84 in synthetic biology and 55 in nanotechnology
  • Some websites couldn’t be parsed
    • For example, Ziptronix Inc. was purchased by Tessera in 2015 [1], and its site no longer exists

Method for Narrative analysis

• Core question: How do the narratives constructed by companies’ differ across sectors?

• Method:
  • Use LDA to Identify topics for each paragraph in each website
  • Map transition probabilities between topics
  • Use these topic and transition mappings to explore the dominant narratives in each sector
Website analysis: modeling narratives

• Understanding narrative through paragraph topics
• One topic model across all sectors together
• Perplexity doesn’t provide much information here
• Number of topics: 28
## Top Topics Descriptions

<table>
<thead>
<tr>
<th>Overall Rank</th>
<th>Topic Concept</th>
<th>Top 3 Associated Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“Research”</td>
<td>product, technology, research</td>
</tr>
<tr>
<td>2</td>
<td>“Product”</td>
<td>product, use, package</td>
</tr>
<tr>
<td>3</td>
<td>“Cell Biology”</td>
<td>system, cell, cancer</td>
</tr>
<tr>
<td>4</td>
<td>“Materials”</td>
<td>material, provide, solution</td>
</tr>
<tr>
<td>5</td>
<td>“Biotech”</td>
<td>mass, cytometric, use</td>
</tr>
</tbody>
</table>

- **Most Common Topic by sector:**
  - **Nanotechnology**: “Product”
  - **Synthetic Biology**: “Research”
  - **Green Technology**: “Cell Biology”
• **Research:** “The FACTORIAL™ assays have been extensively validated over the years of research contract work for biopharmaceutical companies, academia, and regulatory agencies.” [Synbio, www.attagene.com]

• **Product:** “ABBOTT, BIGFOOT PARTNER ON DIABETES CARE
Abbott and Bigfoot Biomedical have entered into an agreement to develop breakthrough diabetes technologies.” [Nanotechnology, www.abbott.com]

• **Cell Biology:** “TECHNOLOGY FOR MEDICAL DIAGNOSTICS
Medical infrared (IR) Imaging, sometimes known as Thermography, offers interesting diagnostics for many diseases, bruises and other surface injuries. It is a technique that can image the temperature distribution, blood flow and other irregularities resulting from various disease related abnormalities...” [Green Goods, www.magnoliaoptical.com]
Topic Transitions

- The ‘most likely’ topic sequences differ across area

<table>
<thead>
<tr>
<th>Area</th>
<th>First Para</th>
<th>Second Para</th>
<th>Third Para</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green Technology</td>
<td>‘company’ (develop, company, product)</td>
<td>‘product/system’ (system, product, substrate)</td>
<td>‘energy’ (electric, research, energy)</td>
</tr>
<tr>
<td>Nano Technology</td>
<td>‘biotech’ (technology, develop, assay)</td>
<td>‘engineering’ (product, learn, engineer)</td>
<td>‘DNA Technology’ (DNA, technology, system)</td>
</tr>
<tr>
<td>Synthetic Biology</td>
<td>‘cell technology’ (mass, cytometric, use)</td>
<td>‘product/system’ (system, product, substrate)</td>
<td>‘engineering’ (product, learn, engineer)</td>
</tr>
</tbody>
</table>
## Topic Transitions cont’d

<table>
<thead>
<tr>
<th>Area</th>
<th>First Para</th>
<th>Second Para</th>
<th>Third Para</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green Goods</td>
<td>‘research’ (product, technology, research)</td>
<td>‘solution’ (system, product, solution)</td>
<td>‘research’ (product, technology, research)</td>
</tr>
<tr>
<td>Nanotechnology</td>
<td>‘DNA Technology’ (DNA, technology, system)</td>
<td>‘technology solution’ (technology, product, system)</td>
<td>‘solution’ (system, product, solution)</td>
</tr>
<tr>
<td>Synthetic Biology</td>
<td>‘engineering’ (product, learn, engineer)</td>
<td>‘industrial’ (industry, technology, product)</td>
<td>‘technical innovation’ (technology, new, advance)</td>
</tr>
</tbody>
</table>
Discussion

• The topical order in which a narrative unfolds reveals the firm or entrepreneur’s approach to building storylines
  • Storylines may be packaged into plots of expected patterns and conclusions (Downing, 2005)
  • Our results suggest a sectoral “dominant logic” of plots appearing in nanotechnology, synthetic biology and green goods, but further investigation is needed

• Why do these narratives matter?
  • Stories package “factual information about [a firm’s] stock of tangible and intangible capital into a simpler, more coherent and meaningful whole” (Martens et al., 2007)
  • Prior research has shown that subjectively defined “symbolic management” activities facilitate resource acquisition and enhanced performance outcomes (Zott and Huy, 2007)
  • Storylines and plots emerge and congeal to create “niches” where technology developers can co-interpret opportunities and marshal resources in networked settings (Geels and Smit, 2000)
Methodological limitations and next steps

- Potential bias introduced when building sample frame (patents) and filtering assignees to create the final sample (via SAM.gov)
- Full probability distribution from topic model not currently used
- Deeper exploration of narrative structure
- Other areas of exploration:
  - Use of image data to help describe firm websites
  - Improving construct validity, e.g., disentangling mentions of “universities” as a way of signaling reputation, disclosing meaningful partnerships, or revealing relevant academic training and skills of staff (c.f., Arora et al., 2016)
Acknowledgements

This research is supported by the National Science Foundation, Award 1646773. The findings in this presentation are those of the authors and do not necessarily reflect the views of the National Science Foundation.
Thank you

Sarah Kelley – skelley@air.org
Sanjay K. Arora – sarora@air.org
Evgeny Klochikhin – eklochikhin@gmail.com
Sarvothaman Madhavan – smadhavan@air.org