Patent Overlay Mapping:
Visualizing Technological Distance

Based on collaborations between
Nils C. Newman¹
Ismael Rafols²
Alan L. Porter³
Jan Youtie⁴
Luciano Kay⁵

¹ IISC, Atlanta
² SPRU, University of Sussex, Brighton
³ School of Public Policy, Georgia Institute Technology, Atlanta
⁴ Enterprise Innovation Institute, Georgia Institute of Technology, Atlanta
⁵ Program in Science, Technology, and Innovation Policy, Georgia Institute of Technology, Atlanta
Context: proliferation of global maps of science

Until 2000s, very few:
- Small et al. (1985), Small (1999), Bassecoulard and Zitt (1999)

In recent years many:
- Moya-Anegon et al. (2004, 2007), Klavans and Boyack (2005, 2009), Balaban and Klein (2006), Rosvall and Bergstrom (2009), Leydesdorff and Rafols (2009), Janssens et al. (2009), Bollen et al. (2009), etc.
Motivation: the shifting structure of science

From ‘The tree of knowledge’
- Hierarchical
- Branching into subdisciplines and specialties
- Matching the social structure of university departments

... to a web of ‘ways of knowing’
- Changing social contract (Gibbons et al. 1994)
- Increasing institutional hybridity (Etkowitz and Leydesdorff, 2000)
- Dissonance between epistemic and social structures
Approaches to mapping the sciences

Local science maps:
- Developed since the 70s (Small 1973)
- Widely used since late 80s
- Improvements in visualization tools in the ’00s

Portraying *local* developments, i.e. *internal* dynamics *within* a field
- Co-citation of articles (Small, 1997)
- Co-word analysis (Callon et al. 1986)
- Co-classification of articles (Noyons and van Ran, 1998)
- Co-citation of journals (van den Besselaar and Leyesdorff, 1996)
Approaches to mapping the sciences

Global Science Maps:
- Small et al. (1985), Garfield (1987)
- But methods seen as unstable and problematic (Leydesdorff, 1987)
- Many choices:
  - Unit of analysis
  - Similarity measures
  - Dimension reduction
  - Visualisation technique and layout
- Expectations: Highly dependent on the specific choices
Surprise: robustness in core structure

Layout can be appear very different, but strongest linkages still similar
Klavans and Boyack (2009)

Moya-Anegon et al. 2007

Boyack, Borner and Klavans, 2009
And the science maps are useful

1. Benchmarking - How similar/different are organisations A & B?

2. Capturing temporal change - How do the patterns evolve over time?
If we can make a map of Science, can we make a map of Patents?
The problems.....

What to use as an equivalent to citation patterns?
  – Are patent citations functionally equivalent to journal citations?

What to use as an equivalent to ISI subject categories?
  – Would IPC’s work as suitable replacement?

How do we get the data?
  – How to get sufficient quantities of data?
The IPC problem.....

If we use IPC’s, what level should be used?
  – Section, Class, Sub Class or Group?

The Sub Class level makes sense but.....
  – The Sub Class level has the population problem

A61K =

A42B =
A potential solution

Fold the IPC category up to create relatively similar sized grouping

- For large population IPC’s, use the smallest Sub Group level

- For small population IPC’s, aggregate up to General Group level, Sub Class or Class

- Establish a floor cut-off and drop very small aggregated populations

<table>
<thead>
<tr>
<th>85709</th>
<th>A61K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>A61K 8/00</td>
</tr>
<tr>
<td>724</td>
<td>A61K 8/02</td>
</tr>
<tr>
<td>1082</td>
<td>A61K 8/04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>85709</th>
<th>A61K</th>
</tr>
</thead>
<tbody>
<tr>
<td>2706</td>
<td>A61K 8/00</td>
</tr>
<tr>
<td>1082</td>
<td>A61K 8/04</td>
</tr>
</tbody>
</table>
The Experiment

To test this approach:

– Pulled the EP Authority population from PatStat for records issued from 2000-2006
  • Source was the Fall 2010 version of PatStat
  • The date range was selected to keep within one IPC version (IPC 7)
  • The set contained over 760,000 records (a mix of grants, drops, etc..)
  • In reality, the search was a record range from EP0968708 (which was published in Jan 2000) and ends with EP1737233

– Collected the instances of IPC’s for all the records

– Each IPC with an instance count greater than 1,000 was kept in its original state

– Each IPC with an instance count less than 1,000 was folded up to the next highest level until the count exceeded 1,000 or the Class level was reached
  • During the folding, any other IPC’s with counts exceeding 1,000 in the same branch were left out of the folding count

– If at the Class level, the population was less than 1,000, the IPC code was dropped for being too small to map
The Experiment continues...

- The cut-off at 1,000 produced 466 IPC entities
  - Several other cut-offs were tested

- The 466 IPC entities were then mapped to thesaurus

- Returning to PatStat the patents cited by the target records were pulled from PatStat and their IPC’s (if available) were mapped the 466 IPC entities
  - No cut-off on Cited IPC’s

- This data collection allowed the creation of table containing
  - Patent Number
  - IPC Number
  - Cited Patent Number
  - Cited IPC Number
The Experiment continues...

- Txt2Pajek was used to put the data table into a Pajek format

- The .net file was transformed by doing
  - Net>Transform>Remove>Multiple Lines> Number Lines

- It was then saved as a .dat file

- Which was used to create a CosSim file using R stats code in Cited and Citing direction

- A factor analysis of the Cosine similarity matrix of cited to citing patents was used to reduce the 466 categories into 35 “macro patent groupings”
  - Factors from 10 to 40 were tested
  - Using 35 factors appeared to have the greatest face validity
  - The 466 categories were color coded based on the factors

- The IPC codes were converted to text labels using a manual process based on the IPC definitions

- The text labels for the categories were used to create a text label for each factor
The result looks like...
But is it useful?

• The following slides contain patent overlays for specific companies and specific technologies based on 466 categories

• Data:
  – Overlays are based on existing datasets of nanotechnology patent applications for Samsung, IBM, Philips, and DuPont and existing datasets of nanotechnology on Nano-Biosensors, Dye-sensitized solar cells, and Graphene
  – Coverage period: 2000-2006

• Graphs:
  – Same scale was sought for all overlays to be able to compare
  – Color of labels not always coincide, a few were modified to match node colors

• Please consider:
  – Alternative companies may be profiled if necessary; we worked with those at the top of the rank in terms of the number of patent applications
  – We used labels for broader categories, we may replace with labels for each node, but the resulting overlays are a bit messy
Issues

- The thesaurus provided to create 466 groups does not match many IPC classes, e.g. with companies.
- Modifications of thesauruses were necessary to be able to match IPC classes in our existing datasets.

<table>
<thead>
<tr>
<th>Company</th>
<th>Total Records</th>
<th>Missing Records</th>
<th>% not shown in overlay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samsung</td>
<td>979</td>
<td>204</td>
<td>21%</td>
</tr>
<tr>
<td>IBM</td>
<td>278</td>
<td>63</td>
<td>23%</td>
</tr>
<tr>
<td>Dupont</td>
<td>172</td>
<td>45</td>
<td>26%</td>
</tr>
<tr>
<td>Philips</td>
<td>147</td>
<td>16</td>
<td>11%</td>
</tr>
</tbody>
</table>

e.g. not matched by thesaurus in corporate datasets:

<table>
<thead>
<tr>
<th>IPC</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>B82B-3/00</td>
<td>289</td>
</tr>
<tr>
<td>B82B-1/00</td>
<td>125</td>
</tr>
<tr>
<td>C09G-1/00</td>
<td>12</td>
</tr>
<tr>
<td>C09G-1/02</td>
<td>12</td>
</tr>
<tr>
<td>G06N-99/00</td>
<td>5</td>
</tr>
<tr>
<td>G12B-21/00</td>
<td>6</td>
</tr>
<tr>
<td>G12B-21/02</td>
<td>6</td>
</tr>
<tr>
<td>C09G-1/04</td>
<td>4</td>
</tr>
<tr>
<td>G01Q-80/00</td>
<td>7</td>
</tr>
<tr>
<td>G06N-3/00</td>
<td>4</td>
</tr>
<tr>
<td>G06N-3/12</td>
<td>4</td>
</tr>
<tr>
<td>G01Q-60/24</td>
<td>6</td>
</tr>
<tr>
<td>G01Q-90/00</td>
<td>3</td>
</tr>
<tr>
<td>G12B-21/08</td>
<td>3</td>
</tr>
<tr>
<td>G01Q-70/00</td>
<td>3</td>
</tr>
<tr>
<td>G01Q-70/12</td>
<td>3</td>
</tr>
</tbody>
</table>
Dye-sensitized solar cells
Conclusions

1. Potential new tool for research management
   - A somewhat intuitive and straightforward method
   - Based
     • on conventional categories
     • the core structure of patents (hopefully)

2. Proposed conditions of application:
   - Large set (>1500 pubs; low resolution >100)
   - Traceability, transparency, parsimony
   - Best contrasted with local maps

3. Potential applications:
   • Benchmarking
   • Exploration of collaboration
   • Capture changes over time
Next Steps

- Update to the current version of PatStat
- Switch to IPC8
- Work with Grants only
- Expand beyond EP?
- Work on the labels
- Segment maps by Year or Year Ranges to look for stability or instability
- Compare results with other global patent mapping efforts
- Experiment with potential connections between Patent Map and Science Maps in technologies with strong science links
- Move beyond IPC’s?
- Your suggestions....
Any Questions?

Thank You!

Nils Newman
newman@iisco.com
www.iisco.com