Extending a co-authorship network analysis to include theses

Lars Vilhuber* Carl Lagoze† Ben Perry‡ others

October 27, 2014

1 Graph components

1.1 Thesis

The data for the thesis components are derived from RePEc Genealogy. It encompasses a single *entity* type (thesis), a single *activity* (implicit in the data: Ph.D.) combined via two *relationships* or *roles* – advisor and author of the thesis – as well as associated *agents*.

1. Ph.D.: Activity
2. Entity: Thesis
3. Agent: Advisor
4. Agent: Ph.D. Candidate/Author

Note that we will define a generic thesis "entity" with each thesis in the database being a specialization of this generic entity.

Agents

```
<prov:agent prov:id="repec:pab175">
  <prov:type>prov:Person</prov:type>
  <foaf:givenName>John M. Abowd</foaf:givenName>
</prov:agent>

<prov:agent prov:id="repec:pze9">
  <prov:type>prov:Person</prov:type>
  <foaf:givenName>Arnold Zellner</foaf:givenName>
</prov:agent>

<prov:agent prov:id="repec:phe22">
  <prov:type>prov:Person</prov:type>
  <foaf:givenName>James M. Heckman</foaf:givenName>
</prov:agent>

<prov:agent prov:id="repec:pvi26">
  <prov:type>prov:Person</prov:type>
```

*Cornell University, corresponding author. This work is funded by NSF Grant 1131848.
†University of Michigan
‡Cornell University

```
    <foaf:givenName>Lars Vilhuber</foaf:givenName>
  </prov:agent>
```

(XML file attached : theses-agents.xml)

Entities

```
  <prov:entity prov:id="exn:thesis">
    <dct:title>A doctoral thesis</dct:title>
  </prov:entity>
  <prov:entity prov:id="exn:thesispab175">
    <dct:title>An Econometric Model of the U.S. Market for Higher Education<
        /dct:title>
    <dct:date>1977</dct:date>
  </prov:entity>
```

(XML file attached : theses-entities.xml)

Note that the thesis could be fleshed out with full bibliographic information, although that information may not be available within the RePEc network. Our definition includes a generic "advisor" activity, although we will not use that, instead highlighting that relationship through a "role."

Activities

```
  <prov:activity prov:id="repec:PhD"/>
  <prov:activity prov:id="repec:PhDpab175"/>
  <prov:activity prov:id="repec:advisor"/>
  <prov:activity prov:id="repec:advisorpab1751"/>
  <prov:activity prov:id="repec:advisorpab1752"/>
```

(XML file attached : theses-activities.xml)

Linking them

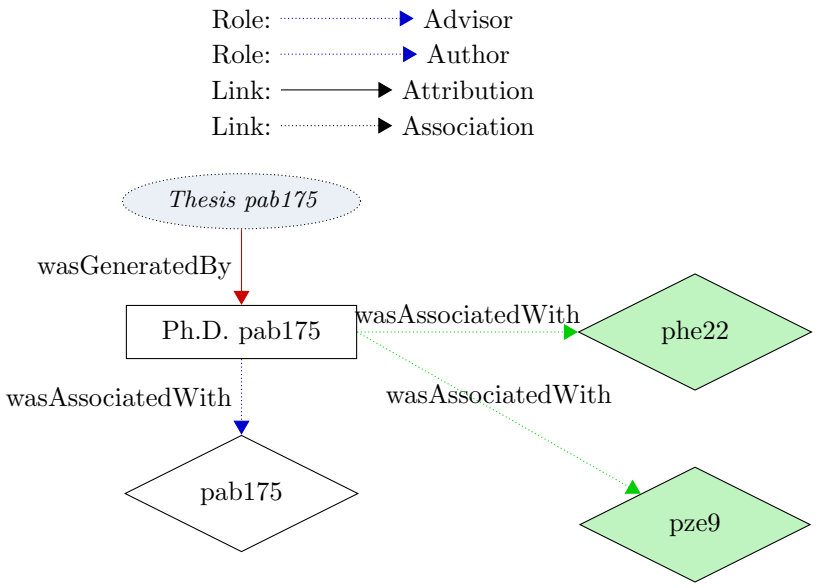
```
  <prov:wasAssociatedWith>
    <prov:activity prov:ref="repec:thesispab175"/>
    <prov:agent prov:ref="repec:pab175"/>
    <prov:role>repec:author</prov:role>
  </prov:wasAssociatedWith>

  <prov:wasAssociatedWith>
    <prov:activity prov:ref="repec:thesispab175"/>
    <prov:agent prov:ref="repec:pze9"/>
    <prov:role>repec:advisor</prov:role>
  </prov:wasAssociatedWith>

  <prov:wasAssociatedWith>
    <prov:activity prov:ref="repec:thesispab175"/>
    <prov:agent prov:ref="repec:phe22"/>
    <prov:role>repec:advisor</prov:role>
  </prov:wasAssociatedWith>

  <prov:wasGeneratedBy>
        <prov:entity   prov:ref="repec:thesispab175"/>
        <prov:activity prov:ref="repec:PhDpab175"/>
  </prov:wasGeneratedBy>
```

Figure 1: Thesis graph



(XML file attached : theses-links1.xml) Pulling these together generates the simple subgraph in Figure 1: However, by explicitly incorporating attributions, we recover the original database construction.

```
<prov:wasAttributedTo>
      <prov:entity    prov:ref="repec:thesispab175"/>
  <prov:agent  prov:ref="repec:pab175"/>
  <prov:role>repec:author</prov:role>
</prov:wasAttributedTo>

<prov:wasAttributedTo>
      <prov:entity    prov:ref="repec:thesispab175"/>
  <prov:agent  prov:ref="repec:pze9"/>
  <prov:role>repec:advisor</prov:role>
</prov:wasAttributedTo>

<prov:wasAttributedTo>
      <prov:entity    prov:ref="repec:thesispab175"/>
  <prov:agent  prov:ref="repec:phe22"/>
  <prov:role>repec:advisor</prov:role>
</prov:wasAttributedTo>
```

XML file attached Incorporating this into the graph, we first obtain a more complex graph (Figure 2): However, collapsing the graph to only the attribution links yields a representation amenable to the usual bipartite graph visualization (Figure 3).

1.2 Co-authorship

The RePEc coauthorship network already exists at `http://collec.repec.org`. (internal representation to be added here).

Implicit in that network is a simple bi-partite network between articles (entities) and authors (agents).

Figure 2: Thesis graph with attribution

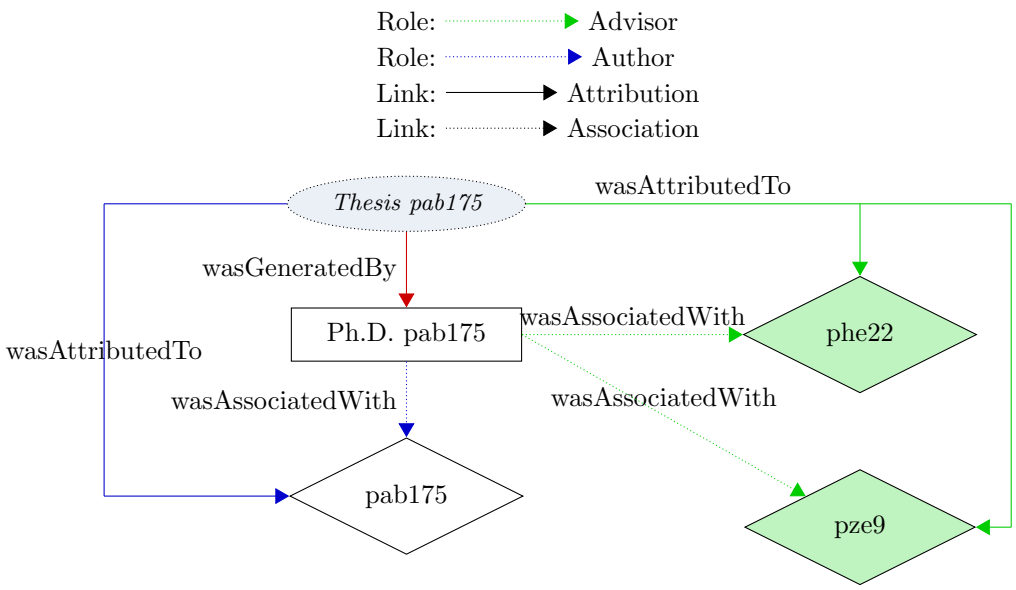
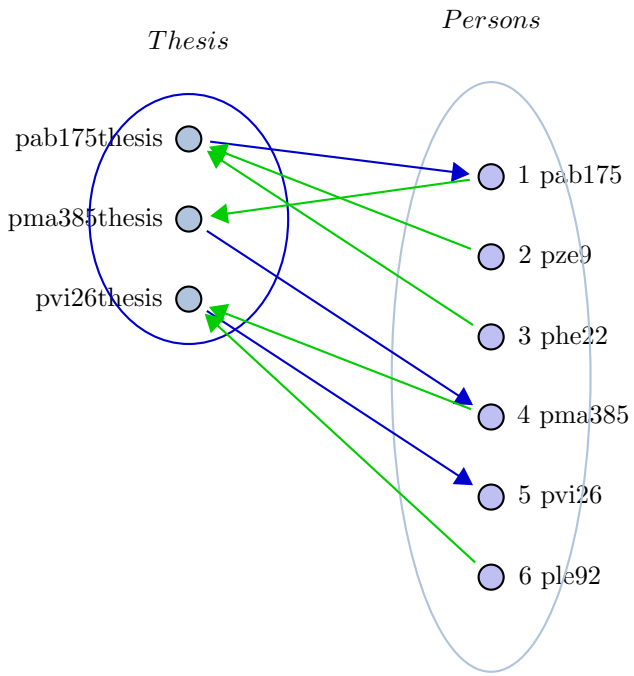


Figure 3: Theses as bipartite graph



Translated into PROV, the definition of agents are as before, whereas the new entities are

Entities

```
<prov:entity prov:id="exn:article">
  <dct:title>A published paper</dct:title>
</prov:entity>
<prov:entity prov:id="hdl:RePEc:eee:econom:v:161:y:2011:i:1:p:82-99">
  <dct:title>National estimates of gross employment and job flows from the
      Quarterly Workforce Indicators with demographic and industry detail
    </dct:title>
  <dct:date>2011</dct:date>
</prov:entity>

<prov:entity prov:id="exn:paper">
  <dct:title>A working paper</dct:title>
</prov:entity>
<prov:entity prov:id="hdl:RePEc:cen:wpaper:10-11">
  <dct:title>National estimates of gross employment and job flows from the
      Quarterly Workforce Indicators with Demographic and Industry Detail
    </dct:title>
  <dct:date>2010</dct:date>
</prov:entity>
```

(XML file attached : coauthor-entities.xml)

Activities

For completeness, we define a research activity to generate articles and papers, although we could directly associate the paper with its authors:

```
<prov:activity prov:id="repec:research"/>
<prov:activity prov:id="repec:research12345"/>
```

(XML file attached : coauthor-activities.xml)

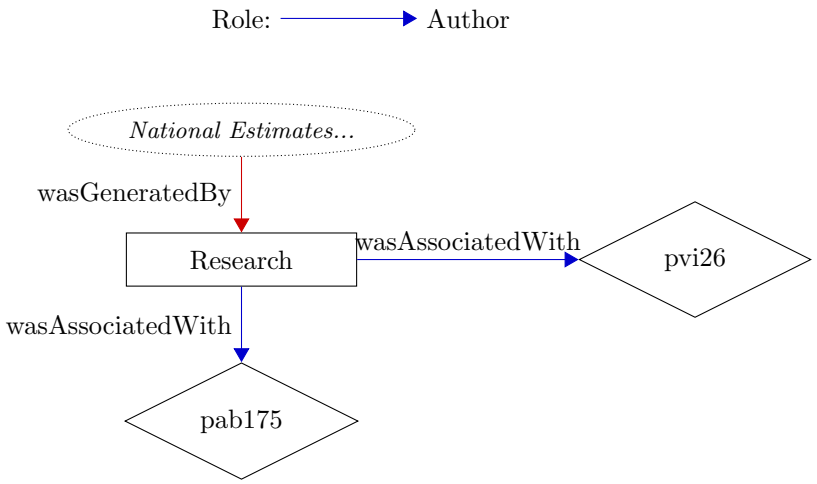
Linking them

```
<prov:wasAssociatedWith>
  <prov:activity prov:ref="repec:research12345"/>
  <prov:agent prov:ref="repec:pab175"/>
  <prov:role>repec:author</prov:role>
</prov:wasAssociatedWith>

<prov:wasAssociatedWith>
  <prov:activity prov:ref="repec:research12345"/>
  <prov:agent prov:ref="repec:pvi26"/>
  <prov:role>repec:author</prov:role>
</prov:wasAssociatedWith>

<prov:wasGeneratedBy>
            <prov:entity prov:id="
                hdl:RePEc:eee:econom:v:161:y:2011:i:1:p:82-99" />
            <prov:activity prov:ref="repec:research12345"/>
```

Figure 4: Authorship with latent research activity




```
</prov:wasGeneratedBy>
<prov:wasGeneratedBy>
        <prov:entity prov:id="hdl:RePEc:cen:wpaper:10-11" />
            <prov:activity prov:ref="repec:research12345"/>
</prov:wasGeneratedBy>
```

(XML file attached : coauthor-links.xml)

Pulling these together generates the subgraph in Figure 4.

The indirect association with the (latent) research activity is implicit in RePEc's linkage of different versions of the same article. Alternatively, these could be noted as revisions; however, the linkage through a research activity is potentially more general. In this case, the attribution is directly coded in the RePEc database, and the implicit research activity is deduced.

```
<prov:wasAttributedTo>
  <prov:entity prov:id="hdl:RePEc:eee:econom:v:161:y:2011:i:1:p:82-99" />
  <prov:agent prov:ref="repec:pab175"/>
  <prov:role>repec:author</prov:role>
</prov:wasAttributedTo>

<prov:wasAttributedTo>
  <prov:entity prov:id="hdl:RePEc:eee:econom:v:161:y:2011:i:1:p:82-99" />
  <prov:agent prov:ref="repec:pvi26"/>
  <prov:role>repec:author</prov:role>
</prov:wasAttributedTo>
```

(XML file attached : coauthor-links2.xml)

For simplicity, a simplified version of the graph, omitting the latent research activity, more closely approximates the bipartite representation:

Figure 5: Authorship with direct attribution

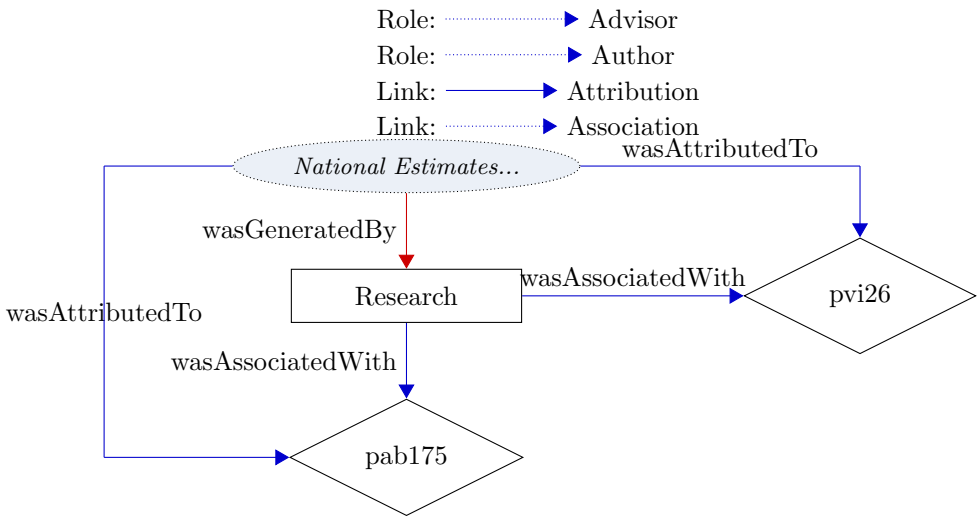


Figure 6: Authorship with direct attribution

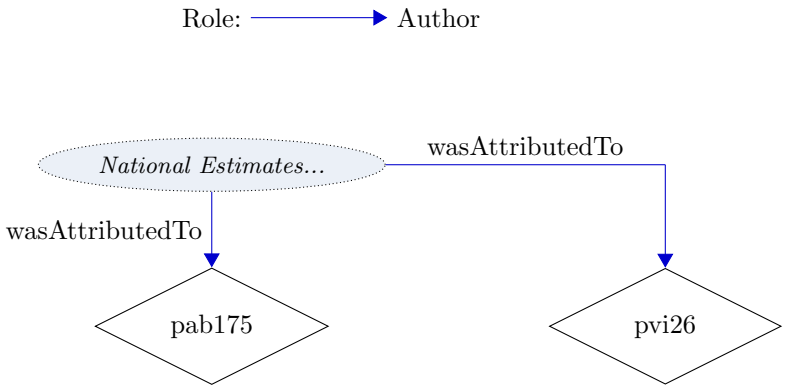
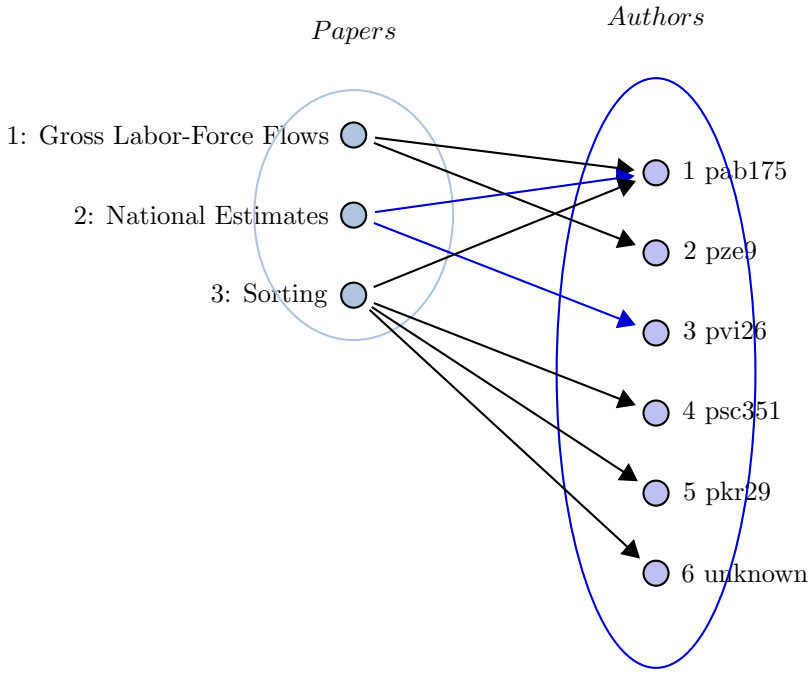


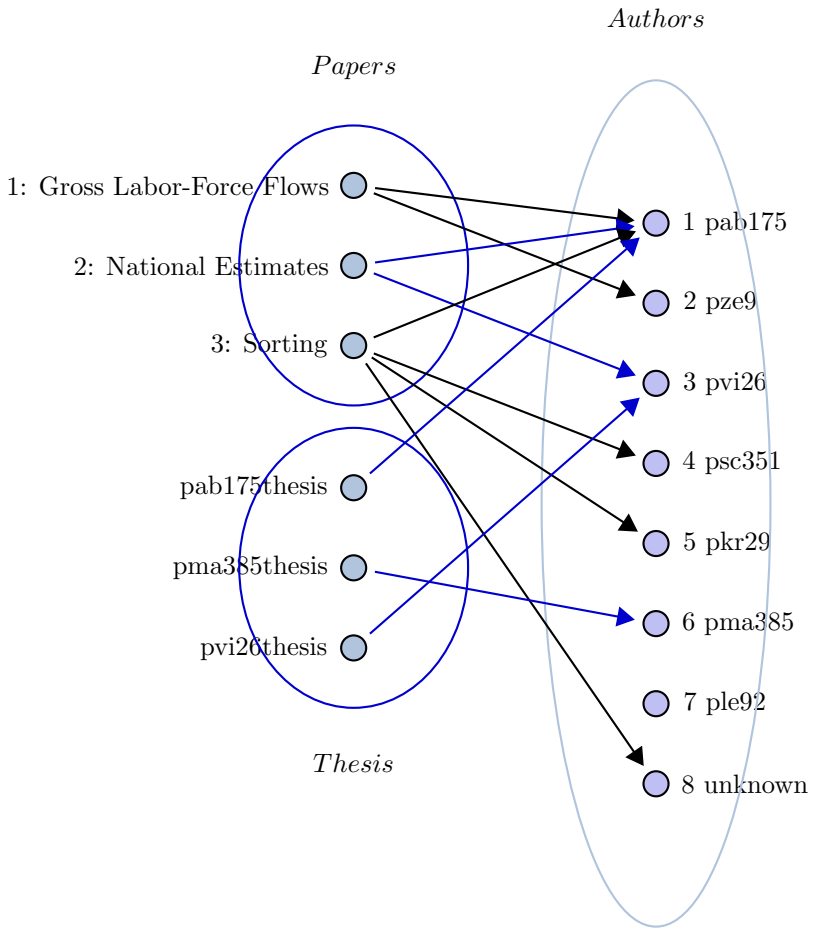
Figure 7: Authorship with direct attribution



2 Combining the subgraphs

We set up the subgraphs such that concepts are identical, and in particular, theses can be treated in much the same way that published articles and working papers can be: they are written artifacts ("entities") of a particular type, that are associated with authors ("agents"). The connectedness of the authorship network is entirely driven by co-author relations: entities with more than one edge generated connected groups. Adding theses to such a network does not change much: almost by definition, each theses has only one author. The average degree of the network might be slightly reduced, in fact, although the RePEc network has an upward bias in the degrees - authors who only publish their thesis and then exit academia, and who those have a low degree, are (probably) under-represented. Figure 8 shows the bipartite representation of the RePEc network with theses added in.[1]

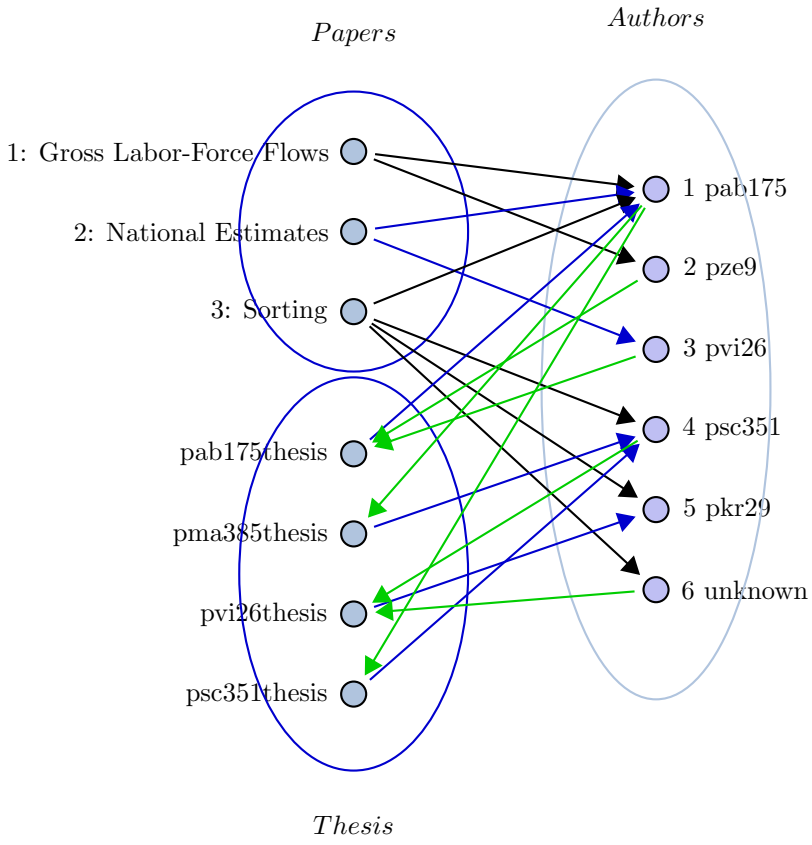
Figure 8: Authorship with theses



However, by adding in a new type of link – advisorship –, the RePEc collaboration network quite naturally is extended, and a substantial amount of new edges are added (Figure 9).

[1]Note that theses are not captured by the RePEc network, and thus don't actually have an entry.

Figure 9: Authorship with theses and advisors



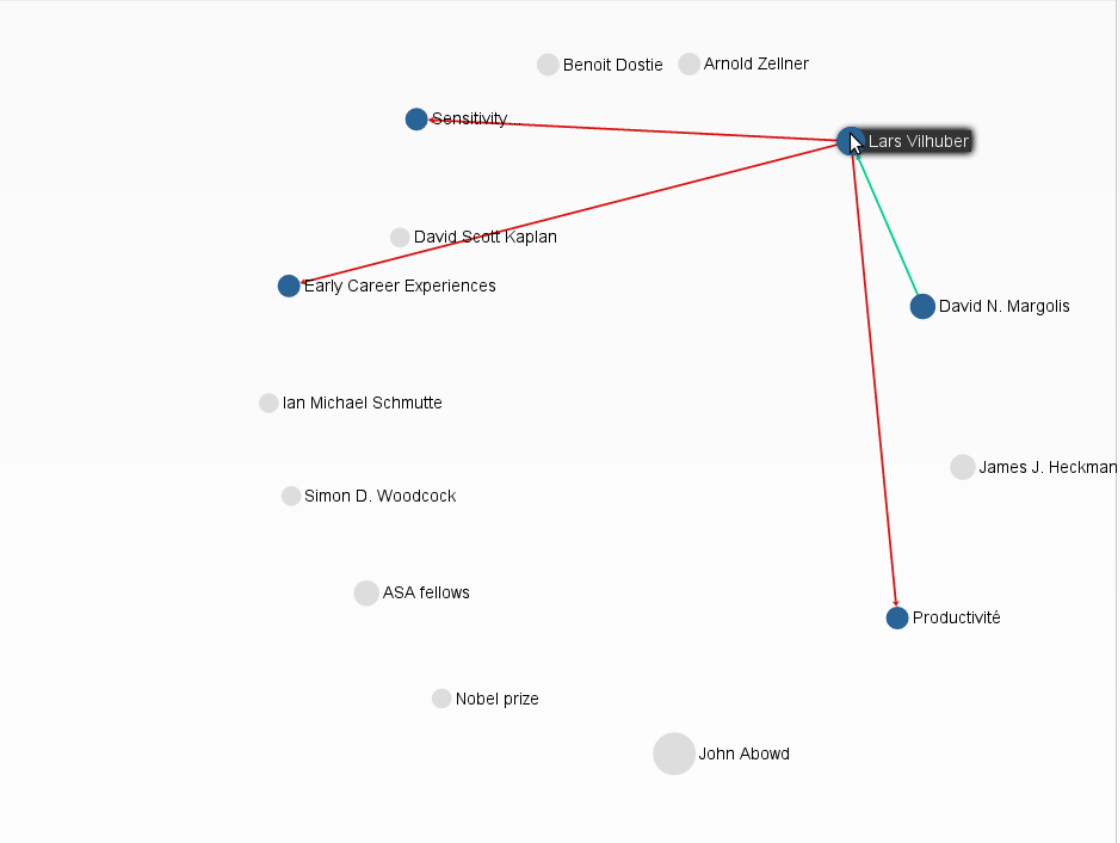
3 Network metrics before and after

We should compute the network metrics before and after the inclusion of thesis advisors in the collaboration network. (average number of edges, average path length, changes in ranking of betweenness).

4 An exploratory interface

Here plug in the graphical interface. See example at `http://www.vrdc.cornell.edu/repecgraph/` and Figure 10.

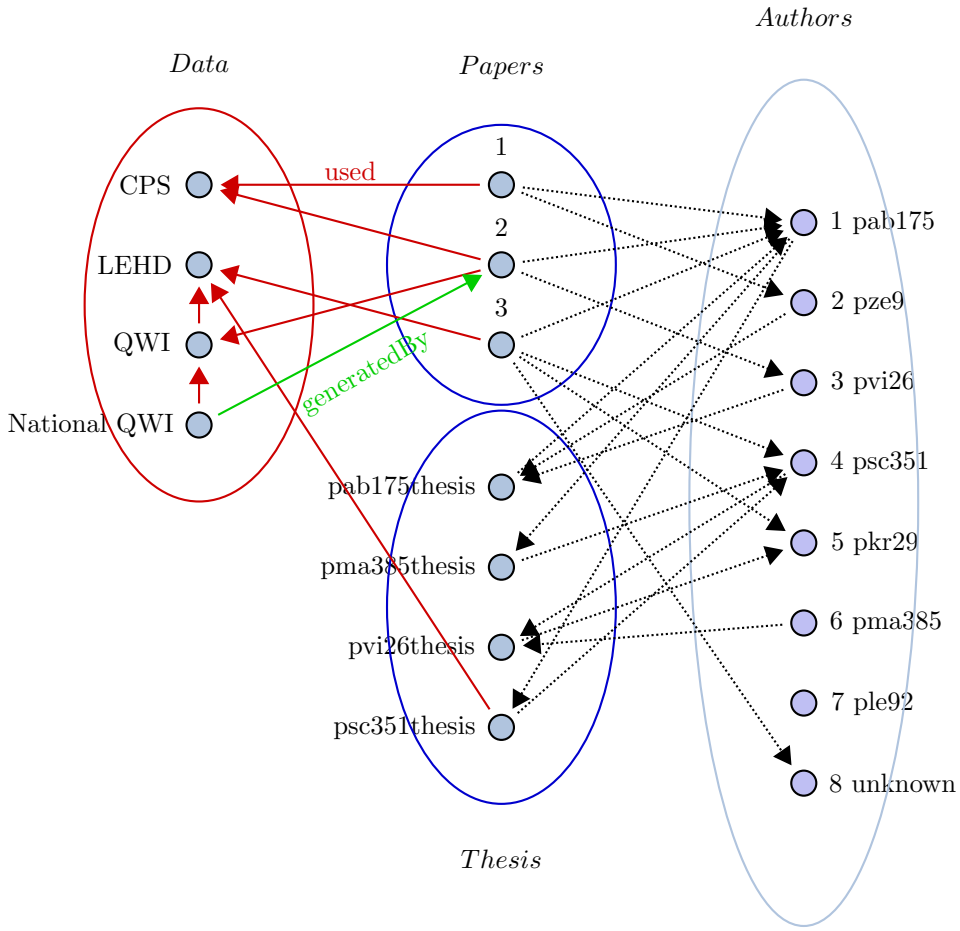
Figure 10: Example webinterface



5 Future work

Already present in the RePEc network are citation links (linking papers among themselves). This work links in with CED^{2}AR work (citations) and other efforts for linking papers and articles to the data used for (empirical) papers. Establishing such links can be represented by a tripartite graph:

Figure 11: Authorship with theses, data



This will ultimately allow to attribute authorship for certain datasets in a clear fashion, which is currently not usual in the social sciences.[2] For an example, see Figure 12.

[2]See ICPSR examples, however.

Figure 12: Authorship of data