An Environment for Building, Exploring and Querying Academic Social Networks

Veselin Ganev, Zhaochen Guo, Diego Serrano, Brendan Tansey, Denilson Barbosa, Eleni Stroulia
Department of Computing Science, University of Alberta
Edmonton, AB, Canada
{vganev,zhaochen,serrano,tansey,denilson,stroulia}@cs.ualberta.ca

ABSTRACT
Social network analysis aims at uncovering and understanding the structures and patterns resulting from social interactions among individuals and organizations engaged in a common activity. Since the early days of the field, networks are modeled as graphs modeling social actors and the relations between them. The field has become very active with the maturity of computational machinery to handle large-scale graphs, and, more recently, the automated gathering of social data. We introduce ReaSoN: a comprehensive set of tools for visualizing and exploring social networks resulting from academic research. In doing so, ReaSoN contributes to the understanding as well as fostering of the social networks underlying academic research. We describe the infrastructure, visualizations and analysis provided in our system, as well as the process of extracting the social networks which are latent in bibliographic and citation databases.

1. INTRODUCTION
The proliferation of web 2.0 REST APIs, through which resources can be shared, and the parallel increased popularity of social-networking applications are creating numerous on-line communities of people with similar interests and objectives. The potential of timely, large-scale on-line collaboration for a variety of purposes and activities is unprecedented. And since, according to James Burke (Connections 1978, BBC), “the easier you can communicate, the faster change happens”, this new technological congruence is bound to have a fundamental impact on the way communities of practice collaborate and make progress on their shared objectives. This is why collaboration practices, and the socio-technical environment in which they occur, are becoming increasingly the subject of study.

Our ReaSoN (REsEarchSOcial Networks) project studies the collaboration of researchers, focusing on computing scientists, with two long-term objectives. First, we examine the co-authorship and citation relations among publications and researchers, in order to understand typical research-collaboration patterns in the field. Second, we are developing tools to support the established collaboration patterns, as well as advance the adoption of new communication and collaboration tools and practices that can evolve and advance the quality and productivity of collaborative research. To that end, we have adopted MediaWiki as the platform on which to collect (programmatically and through user interaction), analyze, display and visualize information on research collaboration. Traditional on-line collaboration tools that make assumptions about explicit organization structures and systematic process workflows do not scale to meet the needs of these varied, loosely coupled communities. More flexible tools are necessary, with less steep learning curves, stronger support for social networking, and capabilities to integrate with other special-purpose tools and resource of interest to distinct subgroups and/or individuals. These requirements are very well met by wikis in general, and MediaWiki in particular. MediaWiki offers an easy-to-learn and easy-to-use platform for users, many of which are already familiar with it through Wikipedia. Second, and more importantly, MediaWiki offers a natural metaphor for understanding the general class of communities-of-practice applications, where resources (corresponding to wiki pages) are shared and collaboratively manipulated by users (who also have their own wiki pages); users can informally discuss resources and other users (through special-purpose discussion pages); and the evolution of the resources, the users and their communications is automatically tracked and reported (in recent-changes and history pages and through email notifications). MediaWiki, as is the case with other wikis, does not afford rich data views, but its software architecture is extensible enough to allow us to alleviate this shortcoming.

We discuss the first version of ReaSoN, providing a test-bed on which to pursue our first research objective, namely the analysis of research-collaboration patterns. In Section 2, we discuss in detail its software architecture. In Section 3, we discuss the notion of network visibility we use to establish the importance of publications, researchers, venues and organizations. Section 4 discusses in detail all visualizations and analysis provided by ReaSoN. Finally, in Section 5 discusses related research. Finally, in Section 6, we conclude and give outline of our plans for the future.

2. ReaSoN INFRASTRUCTURE
ReaSoN is built on top on Annoki, our own extension to MediaWiki. The ReaSoN database is the amalgamation of two large databases of Computer Science literature: DBLP (http://www.informatik.uni-trier.de/~ley/db/index).
2.1 The Annoki Platform

Wikis are collaborative authoring web-based tools, designed to support large-scale on-line collaboration among communities of practice. They are increasingly being adopted by organizations as a means to communicate and inform as well as to collect feedback. MediaWiki is a mature wiki implementation, with several successful web sites built on it (Wikipedia among them). MediaWiki offers several interesting features particularly suited to research collaboration. First, it has special “user” pages, for the personal use of the wiki members, as opposed “regular” wiki pages where content is collaboratively edited (note, that, in principle, a user’s page can be edited by someone other than the user; it is etiquette, rather than technology that prevents this—or rather discourages—from happening). Second, MediaWiki has a “discussion” page for each “regular” page, thus enabling a distinction between “content” and “reviewer’s comments” among the collaborators. Third, MediaWiki supports concurrent editing of pages (with the multiple versions getting merged a-la SVN) and notifications of users when a page, for which they have declared interest, changes; these two features enable a tighter, more synchronous coordination among collaborators. Finally, MediaWiki supports templates, namely organization structures for wiki pages that are defined as their instances. With template support, in addition to free-formatted pages, structured information can be collected.

Namespace-based access control. Each Annoki user has an associated namespace and all pages he creates belong in this namespace. Group namespaces can also be defined to organize wiki pages that “belong” to a group of users. Finally, pages belonging to a “public” namespace are visible to all. In this manner, layers of protection can be supported for personal, project-specific, and publicly accessible content.

Annotations/tags. To enable lightweight cross-referencing of pages, users can annotate pages with their own tags. In this manner, users can superimpose a personal layer of their own on the wiki resources.

Visualizations. Annoki is equipped with two types of rich, interactive, Ajax-based visualizations, wikimap and wiEGOs. Wikimap is a visualization of the whole wiki structure (users, pages, links among pages and edit relations between pages and users). The set of wiEGOs are visualizations of the semantic structures implicit in a set of special template-based pages corresponding to concepts in Blooms taxonomy (i.e., tree, topic, persuasion, brainstorm, story, and decision maps, as well as flowcharts). These visualizations are examples of RICs (Rich Internet Clients) on the resources represented in the wiki pages.

Collaboration and contribution analysis. Extending the default “diff” capability of MediaWiki, Annoki supports analysis of the page edit history at the level of sentences, and collects metrics of each user’s contribution to each page and to the wiki as a whole in terms of sentences added, deleted, and edited.

The namespace and visualizations features motivated our decision to build the ReaSoN prototype on Annoki. In the current ReaSoN prototype all pages are publicly accessible, since all underlying data is publicly available (see Section 2.2). Nevertheless each different type of data (Publications, Researchers, Organizations, Venues, etc.) belong in a corresponding Namespace. In the future however, we will enable access to a private layer where researchers may choose to maintain their own wiki content, in pages belonging to their own namespace. Furthermore, the Annoki wikimap visualization is directly usable in ReaSoN to visually communicate relations among researchers, organizations and publications. More importantly, our experience with embedding RICs in MediaWiki pages was extremely useful in our development of a wide variety of views in ReaSoN (see Section 2.3).

2.2 The ReaSoN Dataset

As mentioned, the ReaSoN database is based on DBLP and the ACM DL. We chose these databases for distinct, and complementary reasons, besides those being the only ones whose maintainers graciously made available to us. The DL has considerably more data for each paper than DBLP, but does not provide unique identification of authors. The DL consists in fact of two databases in one. The ACM DL Guide contains basic bibliographic information about 1.2M publications (as of July 2009) from all of Computer Science, including those of sister societies (such as IEEE). A subset of the Guide, the DL “proper” contains all bibliographic data as well as the full text for approximately 250K publications by ACM only. In terms of citations, the DL contains approximately 6.2M of them. DBLP, on the other hand, is a database of approximately 1M (as of July 2008) BIBTeX records and does not contain many actual citations among papers; however, it is built with a very accurate author disambiguation process (done mostly by its main author).

The publications in the current incarnation of the ReaSoN database (RSN) are essentially the intersection of DBLP and the DL, obtained by matching every publication in the DL to those in DBLP. Besides the citations already given in the DL and DBLP, we also use citations extracted by matching citation strings in the DL (most of which are obtained through OCR and refer to non-ACM publications) against the set of publications in RSN. We keep all metadata available from the DL and from DBLP, including, for each publication, the organizations to which the authors were affiliated at the time of publication. From these, we infer the affiliation history of authors according to their publications. We annotate each affiliation with their corresponding geographical data: country, state/province an city, allowing for some interesting analysis and visualizations. This first version of RSN has:

- publications: 485,267
- authors: 379,188
- citations: 1,301,365
- venues: 3,793 (among which 2,355 are conference series and 623 are journals)
- organizations: 1,865 (among which 1,153 contain “Univ” in their name; also, 336 have no location identified)

From RSN, we build the following networks to analyze and visualize in ReaSoN:

- Publication Citation Network, with 368,064 publications (i.e., nodes) and 1,193,364 citations (i.e., edges);
- Inter-Venue Citation Network, with 2,895 nodes (i.e., conferences or journals) and 135,695 edges (citations).
Table 1: Statistics about the top 1% venues in RSN w.r.t. number of publications.

<table>
<thead>
<tr>
<th>Venue</th>
<th>publ.</th>
<th>cit-out</th>
<th>cit-in</th>
<th>influx</th>
</tr>
</thead>
<tbody>
<tr>
<td>CACM (Comm. of the ACM)</td>
<td>7,311</td>
<td>12,582</td>
<td>39,846</td>
<td>27,264</td>
</tr>
<tr>
<td>TCS (Theoretical Computer Science)</td>
<td>6,536</td>
<td>26,151</td>
<td>13,234</td>
<td>-12,917</td>
</tr>
<tr>
<td>DM (Discrete Mathematics)</td>
<td>5,865</td>
<td>2,348</td>
<td>4,983</td>
<td>2,635</td>
</tr>
<tr>
<td>TC (IEEE Trans. on Computers)</td>
<td>5,033</td>
<td>22,845</td>
<td>21,154</td>
<td>-1,691</td>
</tr>
<tr>
<td>HICSS (Hawaii Conf. on Systems Science)</td>
<td>4,896</td>
<td>27,264</td>
<td>2,906</td>
<td>2,957</td>
</tr>
<tr>
<td>IPL (Inf. Processing Letters)</td>
<td>4,697</td>
<td>6,862</td>
<td>8,448</td>
<td>1,586</td>
</tr>
<tr>
<td>JCT (J. of Combinatorial Theory)</td>
<td>3,652</td>
<td>3,676</td>
<td>4,785</td>
<td>1,109</td>
</tr>
<tr>
<td>WSC (Winter Simulation Conference)</td>
<td>3,648</td>
<td>6,248</td>
<td>3,463</td>
<td>-2,785</td>
</tr>
<tr>
<td>CHI (Comp. Human Interaction)</td>
<td>3,477</td>
<td>9,109</td>
<td>10,420</td>
<td>1,311</td>
</tr>
<tr>
<td>DAC (Design Automation Conference)</td>
<td>3,426</td>
<td>16,285</td>
<td>17,078</td>
<td>793</td>
</tr>
<tr>
<td>ICPR (Int. Conf. on Pattern Recognition)</td>
<td>3,369</td>
<td>80</td>
<td>1,472</td>
<td>1,392</td>
</tr>
<tr>
<td>SIGMOD</td>
<td>3,146</td>
<td>17,272</td>
<td>22,509</td>
<td>1,957</td>
</tr>
<tr>
<td>SAC (ACM Symposium on Applied Computing)</td>
<td>3,069</td>
<td>14,472</td>
<td>1,505</td>
<td>-12,967</td>
</tr>
<tr>
<td>IPPS (Int. Parallel Processing Symposium)</td>
<td>2,929</td>
<td>142</td>
<td>2,689</td>
<td>2,547</td>
</tr>
<tr>
<td>PAMI (IEEE Tran. on Patt. Analysis and Mach. Intelligence)</td>
<td>2,826</td>
<td>18,646</td>
<td>12,672</td>
<td>-5,974</td>
</tr>
<tr>
<td>COMPUTER (IEEE Computer)</td>
<td>2,826</td>
<td>4,087</td>
<td>13,789</td>
<td>9,702</td>
</tr>
<tr>
<td>PR (Pattern Recognition)</td>
<td>2,804</td>
<td>6,649</td>
<td>6,393</td>
<td>-256</td>
</tr>
<tr>
<td>DAM (Discrete Applied Mathematics)</td>
<td>2,773</td>
<td>8,564</td>
<td>2,843</td>
<td>-5,721</td>
</tr>
<tr>
<td>PRL (Pattern Recognition Letters)</td>
<td>2,770</td>
<td>9,966</td>
<td>3,016</td>
<td>-6,950</td>
</tr>
<tr>
<td>CN (Computer Networks)</td>
<td>2,751</td>
<td>9,358</td>
<td>6,664</td>
<td>-2,694</td>
</tr>
<tr>
<td>ISCI (Information Sciences)</td>
<td>2,687</td>
<td>8,310</td>
<td>4,027</td>
<td>-2,583</td>
</tr>
<tr>
<td>TSE (IEEE Trans. on Software Engineering)</td>
<td>2,643</td>
<td>15,258</td>
<td>16,197</td>
<td>939</td>
</tr>
<tr>
<td>VLDB (Very Large Databases)</td>
<td>2,619</td>
<td>26,162</td>
<td>16,864</td>
<td>-9,298</td>
</tr>
<tr>
<td>INFOCOM (IEEE Conf. on Comp. Communications)</td>
<td>2,593</td>
<td>184</td>
<td>15,209</td>
<td>15,025</td>
</tr>
<tr>
<td>CORR (The Computing Research Repository)</td>
<td>2,559</td>
<td>2,896</td>
<td>6,092</td>
<td>3,196</td>
</tr>
<tr>
<td>AAAI (Nat. Conf. on Artificial Intelligence)</td>
<td>2,517</td>
<td>984</td>
<td>7,631</td>
<td>6,647</td>
</tr>
<tr>
<td>JACM (J. of the ACM)</td>
<td>2,434</td>
<td>11,928</td>
<td>20,058</td>
<td>8,130</td>
</tr>
</tbody>
</table>

- Researcher Citation Network: 248,001 nodes (researchers) and 8,322,987 edges (citations derived from the Publication Citation Network);
- Co-authorship Network: 379,188 nodes (researchers) and 2,524,875 edges (joining researchers that have coauthored papers together).

We performed standard statistical tests (e.g., node connectivity analysis) on all networks in RSN. As expected, they all conform with the parameters of typical social networks [3]. Table 1 shows some statistics about the 1% largest venues in the database in terms of the number of publications in them (across all years). As one can see, there is some discrepancy in the coverage of the database, particularly in terms of out-going citations. This happens because the number of citations in the DL varies among venues; naturally, the venues published by ACM have better coverage. Nevertheless, the fraction of venues with high enough citation counts is sufficient to warrant meaningful analysis.

Another interesting observation from the table is that it contains established, highly selective peer-reviewed venues (such as JACM), more accessible flagship magazines (e.g., IEEE Computer), general conferences (e.g., ACM SAC), and premier conferences in specialized fields (e.g., SIGMOD). Perhaps surprisingly, among the notable venues is the non-peer-reviewed CORR.

**Accuracy.** We adapted state-of-the-art approximate matching and duplicate detection techniques [2] to produce RSN. Changes to these techniques are necessary to account for publications with many “copies” (e.g., keynotes in more than one conference) or different versions of the same work (e.g., an expanded version of a conference paper in a journal). We achieve 96% precision (i.e., papers correctly matched) and 88% recall (i.e., papers that should be matched). Most mismatches were caused by the way the DL sometimes subdivides the title of the paper into a “subtitle” or differences in encoding accented characters. These will be addressed in future work. Matching citations strings to publications is much harder as the individual fields (e.g., authors, title, venue, etc.) are not delimited. The same problem occurs when identifying organizations and their locations from the affiliation strings in the DL. We solve both problems using the method in [16], achieving 94% precision and 85% recall for citation extraction and 91% precision and 86% recall for the affiliation extraction.

### 3. VISIBILITY

There are many ways of measuring the importance of nodes in a social network [13, 3]; in ReaSoN we measure reputation through **visibility**: the likelihood of the node “standing out” from the crowd. Node visibility permeates the visualizations and analysis provided in ReaSoN: it is used to rank answers to queries, to prioritize nodes when exploring the network, and to compare venues, organizations, etc. The interpretation of visibility depends on the network: a publication with high visibility will have many highly-visible publications citing it; a highly-visible researcher in the co-authorship network will have co-authored publications with other highly visible researchers.

Our notion of visibility builds on PageRank [11], which we briefly recall. Let \( n \) be a network node with incoming edges from \( p_1, \ldots, p_k \). The PageRank of \( n \), \( PR(n) \) is
\[
PR(n) = d \cdot \left( w_{p_1} \cdot PR(p_1) + \cdots + w_{p_k} \cdot PR(p_k) \right) + \frac{(1-d)}{N},
\]
where \( N \) is the size of the network, \( d \) is the probability of the surfer following an edge (as opposed to “jumping” to a random node), and \( w_{p_k} \) is the preference of the surfer for the edge between \( p_k \) and \( n \) among all edges originating in
The computation is iterative: the PageRank of each node is continuously refined until the process converges (the conditions for this to happen are well understood [8]). In the original Pagerank [11], the surfer follows links randomly with uniformly distributed probabilities, having no prior bias towards nodes in the network. This is modeled by defining \( w_{pi} = \frac{1}{\text{degree}(p_i)} \). However, unlike a random surfer, a trained researcher is highly biased when navigating the network (i.e., following citations). Indeed, the likelihood of a paper being read by a researcher given a citation depends on several factors, chiefly amongst them the visibility of the venue where the cited publication appears. We model the \textit{a priori} bias towards a publication by the visibility of its venue \textit{in the previous year} of the publication. That is, for each year \( y \), we compute the visibility of all publications up \( y \), and accumulate them by venue. Then, for year \( y + 1 \) we weight every incoming citation according to the relative weight of the venue where the paper is published. In the remainder of the paper, we refer to this formulation as the \textit{adjusted} PageRank.

Visibility and Citation Counts. We note that a high citation count is neither necessary nor sufficient for high visibility in ReaSoN. This happens because of the way the modified PageRank is defined: in order for a paper to be visible it must have incoming citations from other visible papers and be published in a venue with other visible papers. To illustrate the difference between citation count and visibility, Figure 1 compares the citation count (node size) and the accumulated visibility (brightness) of the top 1% venues in terms of size. In the Figure, the color of the edges indicate number of citations from few (gray) to many (bright blue). As one can see, some of the most visible venues (e.g., JACM, CORR, AAAI) are not among the largest.

4. THE ReaSoN VIEWS

To discuss the ReaSoN views, let us consider a scenario where a prospective graduate student, interested in “Aspect-Oriented Programming” wants to decide to which University and with which professor to study.

4.1 Publication Page

The student enters the keyword “Aspect-Oriented Programming” in the ReaSoN search textbox and follows the link to the paper with the same name. Figure 2 shows the corresponding publication view. This page reports general information about the publication, including title, publication year, venue, paper keywords, and the URL where more information about the publication can be found, such as the DOI of the publication. In addition, the page includes a graph visualizing how the percentiles for the three adjusted PageRank scores (on the full citation graph in blue,
Figure 4: Publications of an individual researcher.

excluding self citations in green, and excluding citations of co-authors in red) of this paper has evolved over time. The page also shows the paper’s abstract (when available), the publications that the profiled publication cites as well as the publications by which it is cited.

4.2 Researcher Page

The student follows the link to the paper’s first author, accessing the Researcher page, shown in Figure 3. This view lists some basic information about the researcher, including name, URL, an optional description (about), the organizations with which the researcher has been affiliated, with the most recent first.

In addition to this basic information, the piechart in the top right corner of the page summarizes the publication history of the researcher.

- The blue ring wedges correspond to co-authors and the number of papers written by the profiled researcher and each individual co-author; for researchers with a large number of co-authors, only the top 3 collaborators are shown.
- The green ring represents the venues where the researcher has published the most: the top 3 venues and the number of papers that this researcher has published in that venue are shown.
- The orange ring indicates the years during which the researcher been most productive. Each wedge represents a year and the number of papers that the researcher has published during that year.
- The brown ring is divided in 3 percentiles and each wedge represents a visibility bracket and the number of the researcher’s papers that fall into that bracket during the current year (currently, up until 2008).

The researcher page also includes line graphs showing the relative visibility of the researcher over time from five different perspectives. Each line corresponds to the percentile of different measures based on the PageRank scores.

- The blue line shows the percentile of the researcher’s visibility obtained by running the original PageRank on the Author Citation Network.
- The green line shows the percentile of the total adjusted PageRank by all papers of the author on the corresponding year. Two variants of this metric are computed excluding: self citations (brown) and citations by co-authors (orange).
- The red line shows the percentile of the visibility calculated by running the original PageRank on the co-authorship network.

The lines above provide means for richer analysis than what is possible just with citation counts. For instance, observe in Figure 3 that the blue line is below the orange line (for this researcher, the orange line overlaps and covers both the brown and green lines); this can be interpreted as the researcher’s work being cited by a large number of researchers, some of them not being highly influential (which is indicated by the blue line). In other words, the work is influential and visible by most.

In addition, four different h-Index metrics are reported for each researcher, calculated based on the current year as described in [15]. Finally, a tag cloud showing the keywords that have been used by the profiled researcher as well as their relative frequency (font size) and absolute frequency (the number in parenthesis shows how many papers the researcher has written using the particular keyword)

At the top of the page, several tabs link to more information about the researcher. The Papers tab leads to a page containing a sortable table with all the researcher’s papers. Figure 4 shows (partially) the publications of Gregor Kiczales. A similar publications table is provided for the papers Cited-by and Citing the papers authored by Kiczales. The table is interactive, allowing the user to navigate to citing papers and authors.

4.3 Venue Page

Our student, noticing that Kiczales has been publishing frequently in AOSD and ECOOP, decides to learn more about these venues. Each Venue known to ReaSoN also has an associated page that contains information about its Name, Year, URL with more information (if available), and Country and City where the venue was held, in case of conferences. A graph showing the total adjusted PageRank scores of all papers in the venue over time is also included in this page. These are absolute values, not percentiles–thus, the graph shows the percentage of all visibility in the network that is attained by the corresponding venue. (In
passing, the scores plotted in this graph are the ones used for adjusting PageRank as discussed in Section 3).

The venue page also includes a link to a comparison page, which allows the user to select multiple venues and compare their visibility on the same graph. The graph shows the evolution of the visibility of different venues and reveal potentially interesting facts or patterns. Figure 5 compares ECOOP and AOSD for the years 2002 to 2008, showing that AOSD is overcoming the more mature ECOOP. A similar visibility comparison is also available for analyzing the impact of individual publications, mimicking the so-called “test of time” award usually adjudicated by field experts and awarded to influence paper(s) from an earlier instance of the conference. For instance, Figure 6 compares the evolution over time of the 5 most visible papers from the 1997 ECOOP conference—illustrating that while the “Aspect-Oriented Programming” paper has become a classic, it is not the most visible paper from that conference. Moreover, the graph shows that that paper was not cited right away (notice that its first citation occurred only three years after its publication).

4.4 Exploring Keywords

From all the pages that report keywords, i.e., researcher and publication pages, a click on the keyword takes the user to the researchers-using-keyword page, which shows a GIS view of where researchers using the keyword in question are located. Note that only publications that have used the keyword and have a known affiliation are shown here. This means that usually not all researchers using the keyword will be shown on the map. For instance, Figure 7 shows the locations of the known affiliations of all researchers who have published at least one paper with the keyword “Aspect-oriented Programming”. Such visualization quickly reveals biases and preferences in different world regions (e.g., the lack of researchers in Asia publishing on the topic of aspect-oriented programming in this case).

ReaSoN allows the comparison of the frequency with which a keyword appears among the papers of one or more venues over time. In our running example, this analysis would help the student decide which venues to target the submission of new results: one question that arises in such cases is whether to target specialized venues (such as AOSD) or broader, more established venues (such as ICSE, or SIGSOFT).

4.5 Organization Page

ReaSoN visualizes organizations similarly to venues. Each Organization page reports its Name, City, Country, URL and a graph of the organization’s total visibility, i.e., the visibility of all papers published by researchers affiliated with this organization over time. This page also includes a link to a page that compares the visibility of different organizations on the same graph. Furthermore, the page reports the researchers that have been affiliated with the organization, the publications published by these researchers in a sortable table, and a pointer to a GIS view with a marker on the organization’s location. In Figure 8, our student can see a comparison between the Universities of Texas at Austin and of British Columbia, both of which were mentioned in the keywords GIS page of Figure 7, as Organizations with publications in “Aspect-oriented Programming”.

4.6 Querying Social Networks

Besides the pre-defined analysis tools discussed above,
Figure 9: Actors and their relations in ReaSoN.

ReaSoN also provides infrastructure for asking customized queries over the various social networks in RSN and to combine such data with data from external sources. The data model we use builds on the concepts of actors, which are connected through binary relations; both actors and relations may be annotated with descriptive properties. Figure 9 shows the main actors (ellipsis) and relations between them (arrows). Queries in ReaSoN are of the form:

\[
\begin{align*}
\text{SELECT} & \quad \text{MAP}\ |	ext{EXPLORE} \quad a_1, p_1, \ldots, a_j, p_n \\
\text{FROM} & \quad (\text{Actor}) \quad a_1, \ldots, (\text{Actor}) \quad a_k \\
\text{WHERE} & \quad \langle \text{Predicate} \rangle \ |	ext{AND} \ |	ext{AND} \langle \text{Predicate} \rangle
\end{align*}
\]

A \langle \text{Predicate} \rangle is either a relational predicate, that associates a pair of actors, or a selection predicate, that filters out actors based on (the values of) their predicates. For example, our hypothetical student could use the query below to find the co-authors of Gregor Kiczales and their respective publications about “Aspect” published shortly after the classic paper in ECOOP 1997:

\[
\begin{align*}
\text{SELECT} & \quad \text{r1.name, s1.biography} \\
\text{FROM} & \quad \text{scientist s1, researcher r1} \\
\text{WHERE} & \quad \text{sameAsDBpedia(r1, s1) AND r1.name >< 'Gregor'}
\end{align*}
\]

The symbol \(><\) means string containment.

The results of a query can be visualized in tabular form (SELECT queries), explored in the wikimap graph format (EXPLORE queries) or plotted on the world map as in Figure 7 (MAP queries). Figure 10 shows the wikimap exploration view of query above. This view is interactive, allowing one to further explore each node, discovering new connections by expanding existing nodes. Also, one can seamlessly transition back to the “Page” view of every node (Sections 4.1–4.5).

Using External Data. Besides the actors and relations in Figure 9, ReaSoN allows queries using data from external sources (potentially other social networks). Such sources are described using ReaSoN’s actor, relation, and property data model. The data is retrieved through standard APIs and integrated locally in our system. Currently, we have implemented interfaces for extracting social data from FaceBook, as well as more general data from DBpedia and FreeBase. The correspondence between an actor in ReaSoN and its counterparts in an external source is determined based on equality of specific properties. These correspondences are interpreted as special kinds of relations in our model. For instance, the query below asks for the biography in DBpedia of researchers (in ReaSoN) whose names contain “Gregor”:

SELECT r1.name, s1.biography 
FROM scientist s1, researcher r1 
WHERE sameAsDBpedia(r1, s1) AND r1.name >< 'Gregor'

The \text{sameAsDBpedia} predicate matches researchers in ReaSoN to scientists in DBpedia if their names are the same.

5. RELATED WORK

Many Digital Libraries and related services exist today to facilitate the storage and search of bibliographic and citation data (e.g., the ACM Digital Library, IEEE Xplore, DBLP, CiteSeer, Google Scholar, Microsoft’s Libra, etc.). Other online services provide rankings of venues, most notably journals (e.g., ISI’s impact factor, \text{eigenfactor.org}, etc.). Other kinds of tools are beginning to flourish, aimed at helping researchers and educators engage in the development of community portals (e.g., Purdue University’s HubZero). Unlike these tools, ReaSoN is aimed at uncovering (i.e., querying, exploring and visualizing) the complex relations underlying research activities which are latent in bibliographic and citation databases. In doing so, ReaSoN contributes to the understanding as well as fostering of the social networks underlying academic research.

Citation networks are particularly interesting to analyze from a visibility point of view because citations can be interpreted as endorsements of someone’s work. Several methods have been proposed with the goal of identifying the the impact of publication venues. Perhaps the most widely known is the highly controversial [1] Impact Factor [4], defined as the mean number of citations a journal receives in a two-year interval. Other metrics have been proposed aimed at evaluating the prestige of researchers as well; the h-index [5, 14] is one that has been widely used. However, unlike in ReaSoN, such analysis methods focus on quantitative indicators.
better suited to measure productivity, not visibility.

Network Analysis has become very active in recent years, spurred by the widespread availability of social data online [3, 6, 9, 13]. Bibliographic and citations networks have been studied from other angles as well. Newman [10] showed that the research community is a “small world” from the point of view of co-authorship: most researchers are connected through paths shorter than 6), and there is a high “clustering factor” in the network (if researcher r1 works with r2 and r3 works with r4, there is a high chance that r1 and r3 also work together). Pfeiffer et al. [12] studied publications in Genetics and observed a tendency among researchers to work on subjects that have already been worked on quite a bit before; they also observed that researchers tend to cite papers that have already been cited quite a bit (i.e., the rich-getting-richer effect). This observation is in line with the way we adjust PageRank to model visibility in ReaSoN. Landry et al. [7] observed that active collaboration, irrespective of the “intellectual symmetry” of the collaborators or the nature of their affiliations, results in an increase of research productivity.

6. CONCLUSION

In this paper we have described ReaSoN: a comprehensive set of state-of-the-art tools for the extraction, analysis, visualization, and exploration of academic social networks. While focused on the domain of Computing Science, our system is applicable to other domains. We discussed the importance and utility of the various components of reason, the accuracy and coverage of the RSN database we used, and a novel notion of visibility, which better captures the importance of a node in the network than other standard importance metrics. Finally, we discussed a generic data model for describing, querying and integrating social networks, illustrating how it is used to ask queries across multiple data sources.

Our work in ReaSoN is ongoing. We are expanding the portal with means of extracting other relations (namely: service to the community through editorial and program committee participation, and supervisory relationships). We are building a larger version of RSN, also incorporating papers and citations, particularly for non-ACM venues, extracted using Google Scholar. Several open problems remain in the integration and de-duplication of such data. Current techniques employ a batch mode of operation, building the entire database from scratch every time. We are exploring automated ways of continuously gathering information, and integration with more bibliographic data from other sources. Finally, we are using RSN to study the citation behavior in Computing Science, to contrast it with other observed results in other disciplines.

Acknowledgements. This work would not have been possible without the assistance of Michael Ley and the ACM, and financial support from iCORE, Alberta Ingenuity Fund, IBM and NSERC.

7. REFERENCES