Navigating and Annotating Semantically-Enabled Networks of People and Associated Objects

Sheila Kinsella, Andreas Harth, Alexander Troussov,
Mikhail Sogrin, John Judge, Conor Hayes & John G. Breslin

Abstract Social spaces such as blogs, wikis and online social networking sites are enabling the formation of online communities where people are linked to each other through direct profile connections and also through the content items that they are creating, sharing and tagging. As these spaces become bigger and more distributed, more intuitive ways of navigating the associated information become necessary. The Semantic Web aims to link identifiable objects to each other and to textual strings via relationships and attributes respectively, and provides a platform for gathering diverse information from heterogeneous sources and performing operations on such linked data. In this paper, we will demonstrate how this linked semantic data can provide an enhanced view of the activity in a social network, and how the Galaxy tool described in this work can augment objects from social spaces, by highlighting related people and objects, and suggesting relevant sources of knowledge.

1 Introduction

The ability to link to other pages and objects is a key facility of the World Wide Web architecture. It has enabled every web site to become part of a global network of information. More recently, new client server applications such as wikis and blogs have made writing and linking on the Web extremely easy for the average user. The result has been the creation of vast amounts of user-generated content, often organised within online communities. Consequently, there are huge amounts of data becoming available (in real-time or near real-time), creating invit-

ing possibilities for network research, and enabling entirely new avenues for analysis.

In order to take advantage of the huge store of knowledge which is amassing online, we require new methods of navigating this data. The problem is not simply one of countering information overload, although this is certainly pertinent, but of inferring links between relevant sources of information, possibly scattered across several domains. The goal is to enable the user to move through the information space quickly and intuitively by locating relevant related people, concepts and objects at every step.

One problem is that the current link mechanism on the Web does not differentiate between different types of links and does not allow different types of relationships to be expressed. Data is presented as a set of documents and other files, interconnected by hypertext links. The concepts represented in the documents and the types of the relationships between them are not explicitly stated, and can be hard for a computer to infer. Additionally, data accumulated by one user in a particular domain cannot be easily transferred to another domain. For instance, a blogging community may be dispersed over numerous different sites and platforms, and an interest group may share photos on Flickr, bookmarks on del.icio.us, and hold conversations on a discussion forum. A single person may have several separate online accounts, and may have a different network of friends on each. Therefore, the information existing in online social spaces forms massive, intricate and generally disjoint networks of people and objects.

In short, the lack of standards for expressing semantic information in Web 1.0 has resulted in difficulties in aggregation and integration for applications and research, impairing the possibilities for data and network analysis.

Semantic Web research (Berners-Lee et al. 2001) offers the possibility of overcoming these problems by enabling the description of arbitrary objects or concepts, and the relationships between them, using shared machine-readable formats. Semantic data can be viewed as a directed graph where the nodes represent objects or concepts, and the ties represent semantic relationships. A fundamental part of the Semantic Web is the ontology, a data structure specifying the concepts that are needed to understand a domain, and the vocabulary and relationships required to enter into a discourse about it.

Representing Web data in this way allows the expression of different types of relationships between people, between people and concepts or objects, and so forth. Furthermore, these types of relationships are expressed in open formats and can be transferred and understood in the different domains or communities. For

example, the Friend-of-a-Friend (FOAF)[1] vocabulary allows for the expression of the links between people and the things they create and do. The relationships between communities of friends represented in FOAF can be processed in any program that understands the FOAF vocabulary.

There are large and detailed datasets available on the Semantic Web, containing information regarding people, their activity, and their interactions, that are amenable to social network analysis. However, there is a mismatch between two-dimensional graph theory and multi-dimensional social networks (Scott 1988). Real networks contain different types of relations, and are built around objects which connect people together. The use of semantic graphs containing heterogeneous nodes and ties, instead of traditional link-matrices, to represent information about online communities addresses this problem. For example, relation types in an online social network could include "knows" and "sent-email-to" and object types could include publications (linking authors), photographs (linking people depicted in them), and topics (linking those who have an interest in them).

Creating a graph on the Web of different types of objects linked by different types of relationships is a major step towards large-scale computational social network analysis systems that can process various kinds of relationships and objects. However, in order to fully realise the power of these new representation models, users require ways to extract knowledge from the semantic graph and to infer associations between objects that may not be explicitly linked.

In this paper we show how relevant related information can be extracted from Semantic Web data using the Galaxy tool where the output is generated by a spreading activation technique over weighted links. A related method has been applied (Amitay et al. 2004) to derive a geographical focus from a text, based on locations which are mentioned in the text, but that algorithm can operate only on a hierarchical network. Spreading activation has been applied to semantic networks for social network analysis in applications including recommender systems (Liu et al., 2006), community detection (Alani et al. 2003), and modeling trust propagation (Ziegler and Lausen 2005).

To demonstrate our technique, we gather information represented in common formats and represent the data as a semantic graph, consisting of interrelated people, objects and their associated semantic terms. This data is used as input to the Galaxy tool which provides a generic way of ontology-based network mining. We attempt to locate a set of related items within our dataset, given some text re-

[1] http://www.foaf-project.org/

ferring to a particular person or object, or to a set of people and objects. We apply our approach to two example scenarios:

- Ego-centric search, where we attempt to locate a set of nodes closely related to a focus person
- Community detection, where we locate a community centred around two focus people

Our approach makes use of the network of ties existing between people, including not only social connections, but also semantic connections via shared interests or other areas of common ground. The analysis extends further than people and objects that are closely related, to three degrees of separation and beyond.
The main contributions of this paper are as follows:

- We illustrate how a semantic data model of social spaces gives easy access to massive amounts of freely available information
- We describe how Semantic Web data can give improved insights into the activity of a social network
- We present initial results of experiments carried out on a data set extracted from the Semantic Web

2 Object-centered networks

Jyri Engeström, co-founder of the micro-blogging site Jaiku, has theorized that the longevity of social networking sites is proportional to the "object-centered sociality"[2] occurring in these networks, i.e. where people are connecting via items of interest related to their jobs, workplaces, favourite hobbies, etc. On the Web, social connections are formed through the actions of people - via the content they create together, comment on, link to, or for which they use similar annotations.

Adding annotations to items in social networks (e.g., using topic tags, geographical pinpointing, etc.) is an especially useful aid for browsing and locating both interesting items and related people with similar interests. Some popular types of content items include blog entries, videos, and bookmarks. These objects serve as the lodestone for social networks, drawing people back to check for new items and for any updates from those in their network who share their interests. On Flickr, people can look for photos categorized using an interesting

[2] http://www.zengestrom.com/blog/2005/04/why_some_social.html

"tag", or connect to photographers in a specific community of interest. On Upcoming, events are also tagged by interest, and people can connect to friends or like-minded others who are attending social or professional events in their own locality.

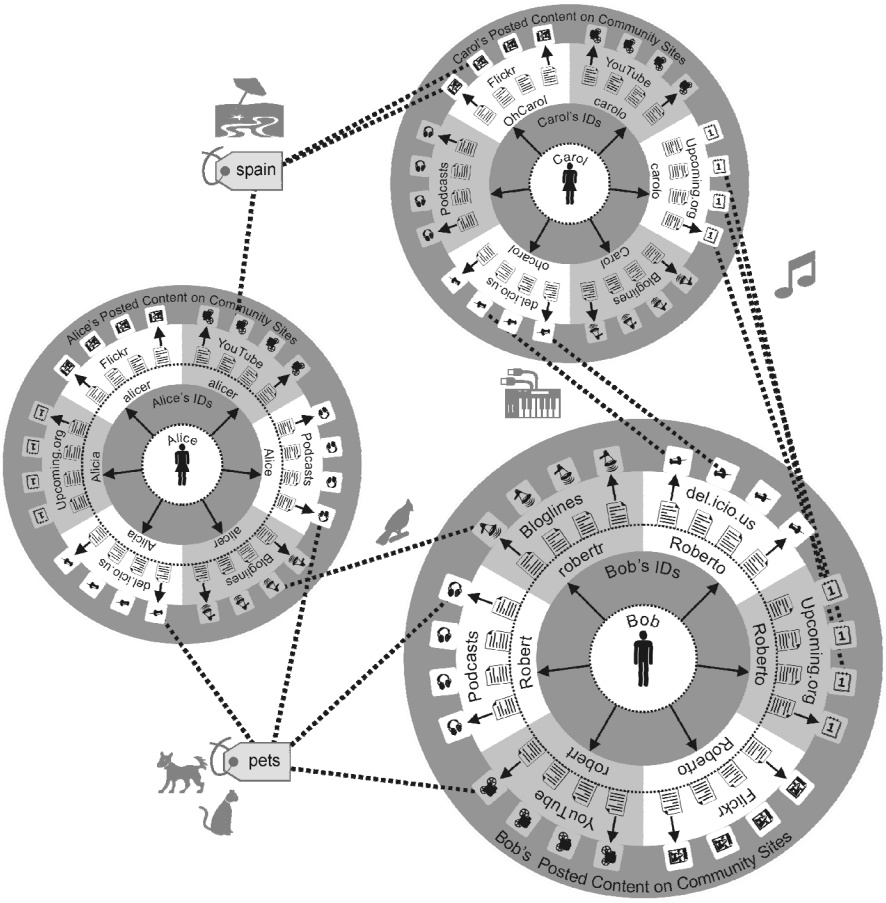

Figure 1: Object-centered social networks are formed by people (using their online accounts) and the content items they act upon

Fig. 1 is illustrative of an object-centered social network for three people, showing their various user accounts on different websites and the things that they create

and do using these accounts. Rather than being connected simply through online social network relationships (i.e. by explicitly-defined friends contacts), these people are bound together through "social objects" of common interest. For example, Bob and Carol are connected through bookmarked websites that they both have annotated on musical keyboards and also through music-related events that they are both attending. Similarly, Alice and Bob are using matching tags on media items about pets and are subscribed to the same blog on birds.

As the connections between people become intertwined with their real-world interests, it is probable that people's social networking methods will move closer towards simulating their real-life social interaction, so that people will meet others through something they have in common, and not by randomly approaching each other.

Since more interesting social networks are being formed around the connections between people and their objects of interest, and as these object-centered social networks grow bigger and more diverse, more intuitive methods of navigating the information contained in these networks have become necessary – both within and across social networking sites (e.g., a community of interest for mountaineering may consist of people and content distributed across photo-, bookmark- and event-centred social networks).

Person- and object-related data can also be gathered from various social networks and linked together using a common representation format. This linked data can provide an enhanced view of individual or community activity in a localized or distributed object-centered social network(s) ("show me all the content that Alice has acted on in the past three months").

The Semantic Web, which aims to link identifiable objects to each other and to textual strings, can be used for linking the diverse information from heterogeneous social networking sites and for performing operations on such linked data. The involvement of objects in social networks on the Semantic Web has been investigated (Kinsella et al. 2007). The Semantic Web is already being used by various efforts to augment the ways in which content can be created, reused and linked by people on social networking and media sites. These efforts include the FOAF project, ontology-enhanced wikis such as the Semantic Media Wiki, the NEPOMUK social semantic desktop[3], and the Semantically-Interlinked Online Communities (SIOC)[4] initiative. In the other direction, object-centered networks can serve as rich data sources for Semantic Web applications. Tim Berners-Lee said in a 2005

[3] http://nepomuk.semanticdesktop.org/

[4] http://sioc-project.org/

podcast, "I think we could have both Semantic Web technology supporting online communities, but at the same time also online communities can support Semantic Web data by being the sources of people voluntarily connecting things together." Users of social networking sites are already creating extensive vocabularies and annotations through "folksonomies" (collections of free-text keywords that are used to tag content items). Since the meaning of these terms is being produced through a consensus of community users, these terms are serving as the objects around which more tightly-connected social networks are centred and formed.

3 Semantic Web

The purpose of the Semantic Web is to enable the online description of arbitrary objects in such a way that software can be used to automatically combine, mine, process, and manipulate data from the Web. Machine-readable descriptions of objects and the relationships between them on the Web enable universal knowledge representation mechanisms on a global scale. For the simplest form of object identification, the same Uniform Resource Identifier is used across multiple sources to reference an object. In many people using the same URI for a particular object, the available data pieces mesh up and form a well-connected and richly-interlinked information space with structured representation features. Layered on top of the foundational URI naming mechanism are a number of other technologies to enable knowledge representation features of increasing sophistication:

- Resource Description Framework (RDF): a universal way of identifying and talking about entities, basic type system (Manola and Miller 2004)
- RDF Schema (RDFS): vocabulary with terms for describing classes and properties, subclass and subproperty relationships (Brickley and Guha 2003)
- Web Ontology Language (OWL): terms for describing classes, inverse properties, cardinality constraints; subset of first order logics (Dean and Schreiber 2004)

Information on the Semantic Web is commonly expressed using the RDF language. An RDF document is composed of a sequence of statements of the form <*subject, predicate, object*>, indicating a directed tie from the subject node to the object node, where the predicate describes the relationship between them.

On the level of RDFS, nodes represent instances of classes, and links represent instances of properties. Classes and the properties which can exist between them are defined in RDFS or OWL. The description of classes and properties form a

vocabulary that can be created or extended as required. For example, vocabularies exist to describe projects, communities, geographical information, and many other domains.

RDF uses the concept of URIs to name all sorts of objects; for example: *http://www.w3.org/People/Berners-Lee/card#i* to denote Tim Berners-Lee, *http://sws.geonames.org/2964180/* to denote the city Galway, *http://deri.ie/* to denote the research institute, and *http://purl.uniprot.orgs/uniprot/Q91474* to denote the protein SHNF1. Objects identified via URIs typically have one or many associated types e.g. *http://xmlns.com/foaf/0.1/Person* or *http://swrc.ontoware.org/ontology#FullProfessor*. The relationships between objects are denoted using URIs, such as the instance-to-type relation *rdf:type*. Namespace prefixes (such as *rdf:*), which indicate the schema to which classes and properties belong, can be used to abbreviate URIs.

In Semantic Web research, the standard way to infer knowledge from a semantic graph is to use an inference engine based on a logic framework such as the OWL to allow logic reasoning on the Web. However, inferring general relationships from graphs can be achieved using techniques other than logic, as we demonstrate in this paper with Galaxy.

4 Dataset

We analyse a dataset consisting of social network information focused around the Semantic Web community. Our model includes people and various related entities. The data under analysis is part of a web crawl of RDF data that was carried out during June/July 2007 using MultiCrawler (Harth et al. 2006). The initial dataset originates from approximately 85,000 sources and consists of over 35 million statements. Object consolidation (Hogan et al. 2007) was performed in order to merge identifiers of equivalent instances occurring across different sources. From the original crawl, we extracted a smaller sub-graph for analysis. The sub-graph is based around the URIs of four people in the Semantic Web community: Tim Berners-Lee, Dan Brickley, Andreas Harth and Tim Finin. We used YARS2 (Harth et al. 2007) to extract all people connected by a path of three or less ties to any of the root nodes, via *foaf:knows* relations. We also included any other nodes connected to these people. The resulting dataset consists of a vast amount of information in many different vocabularies, totalling over 1.2 million statements.

The current version of Galaxy is an early prototype which takes input data expressed in an XML format. However it is planned that RDF support will be

available in the near future. We developed a program to extract specific information from RDF and map it to the required format. For this initial work, we include only a small set of relation types, but it would be possible to extract a much broader range of data. The information we extract is a subset of three vocabularies. Most of the dataset is described using the Friend of a Friend vocabulary (shorthand:*foaf*), which enables the description of people and their relationships with other resources. It also enables the expression of other information relating to a person, such as contact details, as well as publications and other items they have created. Anyone can create their own FOAF file describing themselves and their social network, and social network services can also automatically generate FOAF files for their users, as some, for example LiveJournal, already do. The demand for open, common standards like FOAF is evident from the recent interest in DataPortability[5], an effort by providers of social software, such as Google, Facebook and LinkedIn, to enable users to control and share data across different websites. The information from multiple FOAF files can easily be combined to obtain a higher-level view of the network. We also include some data expressed using the RDF Schema (shorthand:*rdfs*) and Dublin Core (shorthand:*dc*)[6], both of which include properties commonly used to specify the names of resources. There are two main steps to the conversion process - extraction of nodes and ties, and extraction of text labels.

Table 1: FOAF predicates which were extracted and the relation type to which they were mapped

Predicate (foaf:)	Relation type
knows	knows
interest	hasInterest
currentProject, pastProject	hasProject
workInfoHomepage, workplaceHomepage	hasWorkplace
schoolHomepage	hasSchool
made	isMakerOf
maker	madeBy

We derive information from RDF statements based on predicates. All extracted nodes and ties are assigned a type. For instance, all object nodes which occur with the predicate *foaf:interest* are mapped to type ‘interest’. The predicates which we

[5] http://dataportability.org/

[6] http://dublincore.org/

extracted are shown in Table 1, along with the relation type each predicate was mapped to.

Fig. 2 shows the node types which exist in our data model, and the relation types which connect them together. The predicate *foaf:maker* is the inverse of the predicate *foaf:made*. Therefore the corresponding relation type "isMakerOf" is considered to be the inverse of the relation type "madeBy"; in other words, they represent the same relationship, but in opposite directions. None of the other RDF predicates in the data we extracted have an inverse.

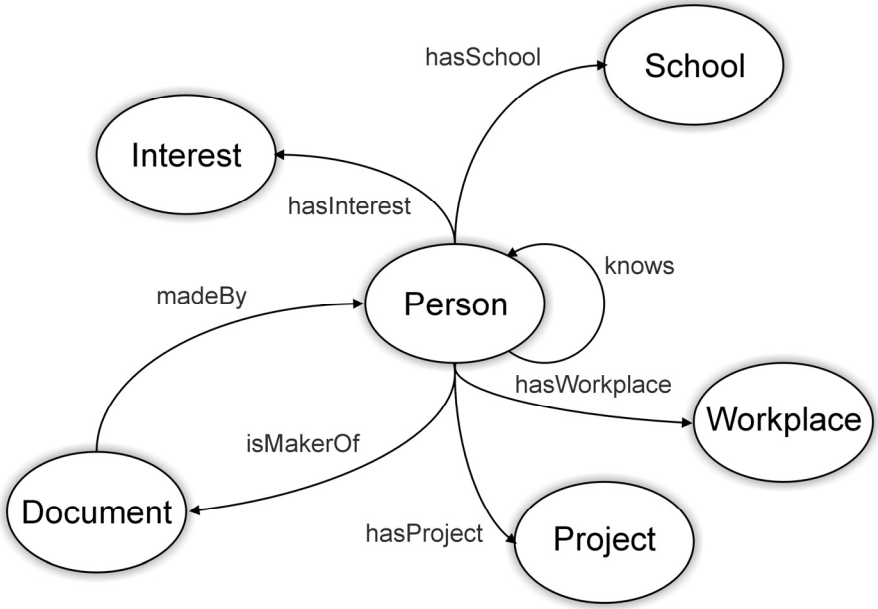

Figure 2: Node and relation types in the data model

We also extract labels for nodes, so that textual references to a particular node will be recognised. For each node type, we made a list of the predicates which indicate that the object node is a name for the subject node. For example, where the subject node is of type Person, this list includes predicates such as *foaf:nick*. Table 2 shows for each node type the predicates which we assume to indicate names. Some nodes may have many different labels. If a node has no name specified, we use the URI of the node as a label.

Our dataset contains 16468 entities and 25028 relationships. Most of the entities are people. The composition of the dataset is shown in Table 3.

Table 2: Node types and the predicates which indicate names

Type	Names
Person	foaf:nick, foaf:name, foaf:firstName, foaf:givenname, foaf:family_name, foaf:surname
Interest	dc:title, dc:subject, rdfs:label
Project	dc:title, dc:subject, rdfs:label
Workplace	dc:title, dc:subject, rdfs:label
School	dc:title, dc:subject, rdfs:label
Document	dc:title, dc:subject, rdfs:label

Table 3 Frequencies of node types in the network

Node Type	Instances	Node Type	Instances
Person	11314(68.7%)	Workplace	443(2.7%)
Interest	2228(13.5%)	Project	339(2.1%)
Document	1956(11.9%)	School	188(1.1%)

5 Galaxy

Galaxy is an ontological network miner designed by the IBM LanguageWare Team[7] for application to tasks in social semantic computing. The Galaxy tool uses a spreading activation algorithm to perform clustering on semantic networks. Instead of the traditional method of hard clustering, which partitions a graph into different groups, Galaxy performs soft clustering, which involves taking a sub-graph based around a set of input nodes, and finding the focus of this sub-graph. The method can be applied to social networks, company organisation charts, or any other set of graph-structured data. Initially, an ontological network of concepts and related terms must be generated based on data provided by the user. Galaxy can then process documents, and identify their main concepts, based on the ontological information. The two main steps to this process are the mapping of terms to concepts, and the location of the main concepts.

[7] http://www.alphaworks.ibm.com/tech/lrw

The Galaxy tool takes a piece of text as input, and then maps terms in the text to concepts in the ontological network. If necessary, the topology of the graph is used in disambiguating terms in the document. The concepts which are identified as corresponding to terms in the text act as input nodes for the spreading activation algorithm. The result of the algorithm is a set of focus nodes, which can be interpreted as those nodes which are most central in the sub-graph based around the input set.

Cognitive psychology and artificial intelligence research model reasoning and memory as processes on neural networks. These networks of neurons and the patterns in which they fire simulate certain aspects of the human brain. There are many different algorithms and implementations which model these processes, one of which, spreading activation (Anderson 1983), is implemented in Galaxy.

In general, the spreading activation algorithm proceeds as follows:

1. Initial activation is set to one or several nodes in the network (e.g. with value 1.0). This initial activation may represent items of interest, context of a document, user profile, etc., and is analogous to sources of light.
2. Activation is spread to neighbouring nodes, but the activation value is normally less then the value of a source. For this, an activation decay parameter is introduced, usually in the range [0…1]. As the activation spreads through the network, different link types may have associated different decay values allowing for different effects like a lower rate of decay through "preferred" links.
3. If activation is spread from a node with many links, those neighbouring nodes will get even less activation to simulate a situation that many similar items get less attention when compared to one unique item.
4. However, if there are multiple paths in the network to some node, its activation will be sum of activations from its inputs. And therefore, it may get activation value even higher than the source.
5. After all activation values are calculated, they are ranked and nodes with higher activation represent important or interesting items or concepts.

Fig. 3 shows how the algorithm finds the focus in a simple linear graph by propagating light of intensity 1 from the nodes at opposite ends of the graph. If the activation is allowed to propagate outwards from the starting points a central node is "illuminated" by both nodes meaning that the level of light is greatest at that point so it is chosen as the focus.

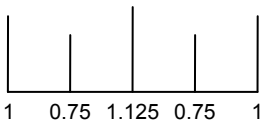

Figure 3: Illustration of the spreading activation algorithm

Galaxy can be used with any kind of graph or tree, and allows for both directed and undirected ties. Various parameters can be tuned to alter the behaviour of the algorithm. This allows a domain-expert to stipulate the properties of a semantic graph that are most important for a particular task. For example, in a graph with different types of relations, some may be considered more relevant than others, depending on the application. The Galaxy tool allows for relation types to be weighted in order to reflect the relative significance of different relationships.

Galaxy can be customised to a range of tasks. Possible application areas include expert-finding, metadata creation, and community detection.

6 Results

In the following we present the results of some sample queries for the semantic graph described in Sect. 4. In each case, we provide Galaxy with a short piece of text, and it uses the topology of the semantic network to extract the most strongly related nodes, based on terms mentioned in the text. Each instance is represented by a URI corresponding to that resource, but here we display human-readable text names. Our queries involve three people: John Breslin, Tim Berners-Lee and Andreas Harth. Firstly, we perform queries for each of these individuals in order to obtain an ego-centric view of their network. Secondly, we perform queries involving pairs of individuals as a means of detecting the community to which they belong.

The objective of the ego-centric queries is to derive an overview of the most relevant available content relating to a particular person. The results for Query 1, “John Breslin”, are shown in Table 4. For this query, Galaxy identifies the people John Breslin and Hannes Gassert, as well as several entities directly related to John Breslin and two entities related to his direct connection Hannes Gassert (Semantic Web at del.icio.us and mediagonal).

Table 4: Results for ego-centric search Query 1: “John Breslin”

Type	Instances
Person	John Breslin
	Hannes Gassert
Interest	Semantic Web at del.icio.us
	Semantic Web
	RDF
Document	John Breslin's blog
Workplace	Semantic Web Cluster, DERI
	DERI
	DERI Galway
	Líon Project, DERI
	Mediagonal
School	National University of Ireland, Galway

The results for Query 2, “Tim Berners-Lee”, are given in Table 5. Galaxy locates the appropriate person and additionally one interest and several documents. uery 3 for “Andreas Harth” locates the person Andreas Harth, one interest and two projects, as shown in Table 6.

Table 5: Results for ego-centric search Query 2: “Tim Berners-Lee”

Type	Instances
Person	Tim Berners-Lee
Interest	Semantic Web
Document	FOAF Document for Tim-Berners Lee
	Tim Berners-Lee's blog
	N3Logic : A Logic For the Web
	Creating a Policy-Aware Web: Discretionary, Rule-Based Access for the World Wide Web
	Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web
	Semantic Web Boot Camp 2007 data

Table 6: Results for ego-centric search Query 3: “Andreas Harth”

Type	Instances
Person	Andreas Harth
Interest	Knowledge Representation
Project	YARS
	SWSE

The objects retrieved by these queries are those which are found to be most relevant to the focus person; not all related entities are shown. The data on which the results are based originates not just from the FOAF files of the individuals involved, but also from other documents which contain references to these people. Results like these could be useful to someone who has come across a reference to these people on the Web and is interested in finding out more related information.

We also experimented with using Galaxy to identify a community, starting with multiple individuals within that community. We chose two queries, each mentioning two people: "John Breslin, Tim Berners-Lee" and "John Breslin, Andreas Harth". The results of these queries are shown in Table 7.

Table 7 Results for community detection Queries 4 and 5

Query	Query 4: “John Breslin, Tim Berners-Lee”	Query 5: "John Breslin, Andreas Harth"
Results	John Breslin	John Breslin
	Tim Berners-Lee	Andreas Harth
	Dan Brickley	Hannes Gassert
	Eric Miller	Aidan Hogan
	James Hendler	Matteo Magni
	Henry Story	Fergal Monaghan
	Charles McCathieNevile	Sheila Kinsella
	-	Siegfried Handschuh
	-	Axel Polleres
	-	Knud Möller

The subjects of our first community detection query, John Breslin and Tim Berners-Lee, are both involved in Semantic Web research. However they are not directly connected to each other. The results show that Galaxy has identified a set of individuals who are located around the two subjects in our query, resulting in a broad view of the Semantic Web community. These people were not identified as

relevant to either of our initial separate queries for John Breslin and Tim Berners-Lee, however when we take the two people together they are found to be important. This is because the activation spreading from both of these nodes overlaps at the nodes in between and raises their rank in the results. These results are based on data aggregated from Tim Berners-Lee's FOAF file, John Breslin's FOAF file, and other documents. This overview of the network is not possible without considering information from multiple sources in our dataset.

The second community detection query involves John Breslin and Andreas Harth. In this query the two people are again Semantic Web researchers, however in this case they work together within the same research institute. The second query therefore has a much narrower focus than the first. All of the people identified by Galaxy for the query "John Breslin, Andreas Harth" are members or former members of the Digital Enterprise Research Institute, and are closely connected to one or both subjects of the query. Most of them were not identified in Queries 1 or 3, because the connection to the focus node was not rated as strongly as, for example, documents authored by the focus node. However in the community detection queries there are now two focus nodes, and the people in the results set are included because they are related to both, which increases the activation of these nodes. As for the previous query, the results are enabled by the aggregation of social networks expressed in multiple interconnected FOAF files.

Although all of the queries given above are very simple, longer text documents can be analysed with Galaxy, for example e-mails and blog posts.

7 Discussion

The examples we have shown in this paper indicate that mining the graph of Semantic Web data using a spreading activation approach allows for the discovery of new relationships between nodes. Evaluating the results returned by Galaxy a more objective way will be a difficult task. This is due to a number of factors. The most common evaluation approaches for recommender-type systems are performed offline using techniques from machine learning and information retrieval such as cross validation and measures of recall/precision (Hayes et al. 2002). In order to conduct such an analysis we require a data set (an ontology), a number of queries, and relevance judgements for those queries on the data set. As a result of difficulties arising from these requirements we have been unable to provide an extensive qualitative analysis here.

Queries are relatively easy to create using use cases and scenarios, however, it should be noted that depending on the user or the task the same query might anticipate different results.

The data available to us is useful for proof of concept testing, but contains a lot of noise and much manual intervention was required to make the subset used in these experiments usable. Due to the novelty of our component's implementation there exists no external standard corpus or dataset (to our knowledge) which is suitable for evaluating this kind of functionality and the cost of manually creating a sufficiently large dataset is prohibitively high.

Relevance judgements for queries require a lot of manual work and investigation, are very subjective depending on who decides what is relevant, and it is very difficult to say if the process has been exhaustive on datasets large enough to be considered suitable for meaningful evaluations.

We see our work as a first step towards using rich web data to gain improved and timely insight into the formation and evolution of social networks.

8 Conclusions

This paper shows how to aggregate and integrate social network information from multiple online sources. We have demonstrated that Semantic Web technologies allow for the collection of real-world data under liberal licenses at an unprecedented scale and at a low cost. We illustrated the benefit of Semantic Web data for social network analysis using the Galaxy tool, which generates a set of related items by a spreading activation technique over weighted ties. We began with an outline of object-centered networks, and described how a semantic data model of social spaces can give an improved insight into the activity of a social network. We then explained the capabilities of the Galaxy tool in ontology-based mining of social semantic networks, and demonstrated how it can provide an enhanced view of networked data. Finally, we presented initial results of experiments carried out on a data set extracted from the Semantic Web, which make use of the network of ties existing between people, including not only social connections, but also semantic connections via shared interests or other areas of common ground. The analysis extends further than people and objects that are closely related, to three degrees of separation and beyond. There are challenges in evaluating the output of systems using web data, and the usage of personal details, even those that are publicly accessible, may create privacy concerns. However we believe that the applications and types of analysis made possible by the free availability of massive

amounts of social information will also give social network researchers a chance to work with huge amounts of real-world data and potentially gain insights into how social networks, both online and offline, form and evolve.

References

Alani, H., Dasmahapatra, S., O'Hara, K., Shadbolt, N. (2003). Identifying Communities of Practice through Ontology Network Analysis. *IEEE Intelligent Systems*, 18, 18-25.

Amitay, E., Har'El, N., Sivan, R., Soffer, A. (2004). Web-a-where: geotagging web content. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. NY, USA.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22, 261-295.

Berners-Lee, T., Hendler, J., Lassila, O. (2001). The semantic web, *Scientific American*, 284, 28-37.

Brickley, D., Guha, R. V. (2003). RDF Vocabulary Description Language 1.0: RDF Schema. *W3C Working Draft*.

Dean, M., Schreiber, G. (2004). *OWL Web Ontology Language Reference*.

Harth, A., Umbrich, J., Decker, S. (2006). MultiCrawler: A Pipelined Architecture for Crawling and Indexing Semantic Web Data. *Proceedings of the 5th International Semantic Web Conference*. Athens, GA, USA.

Harth, A., Umbrich, J., Hogan, A., Decker, S. (2007). YARS2: A Federated Repository for Searching and Querying Graph-Structured Data. *Proceedings of the 6th International Semantic Web Conference*. Busan, Korea.

Hayes, C., Massa, P., Avesani, P., Cunningham, P. (2002). An on-line evaluation framework for recommender systems. *Proceedings of the Workshop on Recommendation and Personalization in eCommerce at the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems*. Malaga, Spain.

Hogan, A., Harth, A., Decker, S. (2007). Performing Object Consolidation on the Semantic Web Data Graph. *Proceedings of the I3: Identity, Identifiers, Identification Workshop at the 16th International World Wide Web Conference*. Banff, Alberta, Canada.

Kinsella, S., Harth, A., Breslin, J. G. (2007). Network Analysis of Semantic Connections in Heterogeneous Social Spaces. *Proceedings of the UK Social Network Conference*. London, United Kingdom.

Liu, H., Maes, P., Davenport, G. (2006). Unraveling the Taste Fabric of Social Networks. *International Journal on Semantic Web and Information Systems*, 2, 42-71.

Manola, F., Miller, E. (2004). *RDF Primer*.

Scott, J. (1988). Trend Report: Social Network Analysis. *Sociology*, 22, 109-2

Ziegler, C. N., Lausen, G. (2005). Propagation Models for Trust and Distrust in Social Networks. *Information Systems Frontiers*, 7, 337-358.