Computing and Informatics, Vol. 33, 2014, 479–502

SEMANTIC SOCIAL NETWORK ANALYSIS FOR AN ENTERPRISE

Charalampos Chelmis, Hao Wu

Department of Computer Science
University of Southern California
3740 McClintock Avenue, Los Angeles, CA 90089-2562
e-mail: {chelmis, hwu732}@usc.edu

Vikram Sorathia, Viktor K. Prasanna

Ming Hsieh Department of Electrical Engineering
University of Southern California
3740 McClintock Avenue, Los Angeles, CA 90089-2562
e-mail: {vsorathi, prasanna}@usc.edu

Abstract. Business processes are generally fixed and enforced strictly, as reflected by the static nature of underlying software systems and datasets. However, internal and external situations, organizational changes and various other factors trigger dynamism, which is reflected in the form of issues, complains, Q & A, opinions, reviews, etc., over a plethora of communication channels, such as email, chat, discussion forums, and internal social network. Careful and timely analysis and processing of such channels may lead to early detection of emerging trends, critical issues, opportunities, topics of interests, contributors, experts, etc. Social network analytics have been successfully applied in general purpose, online social network platforms, like Facebook and Twitter. However, in order for such techniques to be useful in business context, it is mandatory to integrate them with underlying business systems, processes and practices. Such integration problem is increasingly recognized as Big Data problem. We argue that Semantic Web technology applied with social network analytics can solve enterprise knowledge management, while achieving integration.

Keywords: Social networking analysis, Semantic Web, semantic social network analysis, information integration, knowledge networks, cross-enterprise collabora-

tion, corporate knowledge management, collaboration analytics, communication channels, social search, knowledge representation, knowledge acquisition, collective knowledge, collective intelligence, user modeling, ontology engineering

1 INTRODUCTION

An enterprise's success is bounded by its capacity for quick adaptation and spontaneity in technological changes and paradigm shifts, rapid response to customer demands through anticipation of future opportunities, and ability to predict, detect and alleviate risk factors and threats. "*The competitiveness of firms is related to the adequacy of their decisions, which depends heavily on the quality of available information and their ability to capitalize, enrich and distribute this relevant information to people who will make the right decisions at the right moment*" [1].

In modern enterprise, engineers typically spend 40–60 % of their time seeking information [2, 3]. A system that enables quick expert identification and facilitates interdisciplinary cooperations that span organizational charts, lessening time spent on searching for solutions, is pivotal for its success. In high risk operation environments such as smart oilfields [4], shortening the response time required in a failure event may result in excessive environmental and economical savings.

The end-goal for an enterprise is not storing and managing lots of raw data, but instead, to get to newer actionable business insights faster. We argue that in order for enterprises to get to such insights faster, there is an imperative need for a platform that enables quick, rich, and novel data exploration in multiple, intuitive ways, gleaning information from multiple communication mediums and leveraging it into knowledge. However, due to the way data is generated in a modern enterprise, data management has become increasingly challenging.

1.1 Social Interactions: Big Data – Big Opportunities

Knowledge is generated, captured, utilized and shared without being limited to a specific language or system, but encoded in multiple formats, and distributed over various repositories. In large organizations, knowledge can be handled in the form of standard operating procedures, questioning and answering forums, FAQs, internal websites, social network, personal e-mail correspondence, and other means of communication. In this context, knowledge is highly dynamic and constantly evolving, and unless otherwise captured, it becomes "buried knowledge" [5].

Due to the richness and variability of systems and tools available in the enterprise information ecosystem, multiple communication channels between employees have become available. User activity and behavioral data in this context contains valuable information. User's interests in personal and professional level can be discovered, whereas interesting communication motifs can be mined out, enhancing our understanding on employees' communication patterns as well as patterns of information propagation and browsing in enterprise networks.

1.1.1 Sources

Enterprises have been mainly relying on e-mail traffic to share information among coworkers [6]. Analysis of enterprise communication networks [7, 8, 29, 31] has broadened our understanding of information flow in the enterprise. [9] argued that "*information extracted from e-mails could prove useful in a knowledge management perspective*", as it would facilitate expert and community identification. Media like SharePoint and Office Communicator are heavily utilized as part of question-answering and problem solving processes, while Active Directory provides a formal structure for employees to comprehend and navigate through the organizational hierarchy, accommodating their need to identify potential collaborators, research teams and business units around the globe, as well as to discover "interesting" projects that others are currently working on.

The wealth of information available in the context of enterprises, however, is not limited to formal interactions and silos containing structured data. As social media have become phenomenally popular, enterprises have adopted light-weight tools such as on-line forums and microblogging services for internal communication. Employees have been using social network sites and microblogging services to stay in touch with close colleagues or to reach out to employees they do not know, to connect on personal level or to establish strong professional relationships in order to advance their career within the company [10]. Others perceive the use of such services as extra source of company news and events, a mean to promote their ideas or to contribute to conversations revolving around company matters.

Enterprise social interactions analysis may lead to various insights, both at atomic (micro) and collective (macro) level. Meaningful micro analysis could revolutionize employees perception of the working environment, offering them better tools for communication, search and productivity, whereas macro analysis could be used for strategic decision making and informed planning.

1.1.2 Macro Analysis

Enterprises can utilize the results stemming out of informal interactions analysis, to better understand how their employees work together to complete tasks or produce innovative ideas, reveal trends, identify experts and influential individuals, so as to evaluate and adjust their management strategy, team building and resource allocation policies.

1.1.3 Micro Analysis

Similarly, employees can benefit in multiple ways. Recommendation services can provide better results in terms of "interesting" people to connect to, as well as suggest "interesting" discussions for employees to contribute to or projects to get involved in. Information filtering algorithms can better promote subset of news

instead of directly delivering all sorts of irrelevant data to employees, alleviating information overload from them, and enabling them to focus on information that does matter. Information acquisition, such as search for people, data and answers to problems can be significantly sped up, resulting in increased productivity through collaboration and problem deduplication.

1.2 Social Big Data Sources: The Big Mess

In this work, we primarily focus on capturing employees' interests and areas of expertise, as well as mining interconnections between employees' work-related activities and their social interactions on collaboration platforms used in working environments. In practice, users' activities are scattered across various collaboration tools used in the enterprise, leaving behind structured, unstructured or semistructured information traces in multiple formats. A user might choose chat or microblogging services for casual Q & A sessions, e-mail correspondence for document and ideas sharing, and SharePoint for project tracking purposes. Furthermore, a user may adopt different tools for different projects, or utilize different tools for the same project, depending on current needs. In general, the existence of multiple communication channels scatters information related to a specific employee, establishing the need for an integrated view of users' activities across platforms.

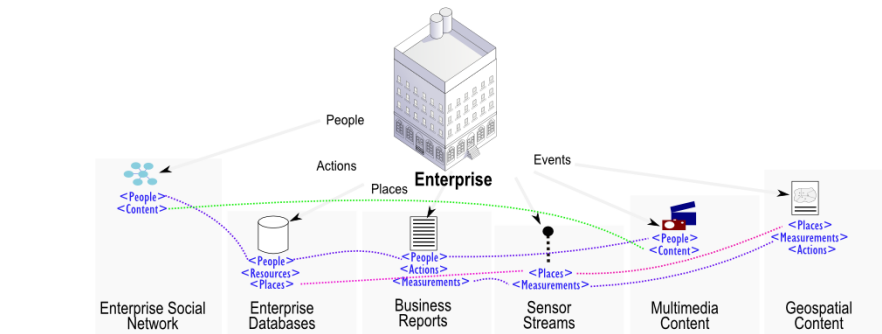

Figure 1. Big data in enterprise

High-volume activity on enterprise communication channels offers unique opportunities for content analysis. Active participation of employees and sharing in informal conversations makes it possible to identify knowledge and expertise by analyzing users' contributed content and its impact to the community. Figure 1 depicts information sources typically found in modern enterprise, each offering unique

business perspectives, highlighting differentiations in underlying formats, update frequencies, and scope.

Integrating users' activities across multiple collaboration tools is not an easy task. Content from collaboration activities may be significantly short (e.g. 140 characters in Twitter) and inherently noisy. For instance, microblogging content does not adhere to any grammatical or syntactical rules, contains slang terms, user-defined hashtags and emoticons or other special characters, which denote emotions or user-defined notions, the semantics of which may be unknown or not previously modeled. Second, users activities on various collaboration tools signal different kinds of relations, personal or professional [10], of unequal importance [11]. Third, information heterogeneity due to different format and schemata or storing mechanism impose further restrictions. Fourth, capturing employees' interests and areas of expertise is a time sensitive task. Existing methods model users' interests based on static profiles or by keeping track of users' collaborative activities. However, users' profiles may be completely unavailable or extremely scarce since users do not often populate enough information to describe themselves. Profiles may be obsolete if users do not constantly update them to match their most up to date interests.

1.3 Summary of Contributions

- We design and build a model that accurately captures the multidimensional nature of complex, informal, social interactions in the form of orthogonal dimensions. Our model encompasses static enterprise characteristics, as well as dynamic collaboration modalities.
- We utilize semantic web techniques for our conceptual modeling and representation. This approach enables seamless integration of shared domain ontologies and linked open data.
- We show that our model facilitates integration, capturing, search, and retrieval of dynamic and constantly evolving enterprise knowledge captured from informal, social interactions.
- We show how our model enhances collaborative data analysis in the enterprise, revealing latent topics, expertise, and interests, both at micro and macro level.
- Our approach leverages knowledge stemming out of informal communication from multiple sources, driving multiple applications, such as team building, recipient recommendation, and event recommendation. We present a case study on a large scale dataset from a Fortune 500 company.

2 REPRESENTATION AND ANALYSIS OF BIG SOCIAL DATA

Researchers have modeled, captured and analyzed interactions between people in a plethora of situations [12]. To better mine and understand such complex interactions, their properties and characteristics, the need for some appropriate representation has emerged.

Researchers have applied graph theory [13] on various graph representations so as to unravel network features, identify the most important actors and discover community structures. Social networks have been represented as *sociograms* [14], in which nodes represent users and arcs represent explicit relationships between them. In order to exploit implicit relationships between users, *tripartite graph* models, have also been proposed [15, 16].

Social networks evolve when users "friend" each other. However, "friend-of" links fail to capture the strength of association between users and explicit relationships between them. For example, two users may be computer programmers, but interested in PHP and Java, respectively. In this scenario, linking users based on a specific programming language misses the latent relationship in the dimension of computer programming. Typically, social networks capture relationships in a one-dimensional manner: two users are connected by an edge carrying the generic "friend-of" label.

Rich human interactions and socially generated data cannot be represented using graph models alone. In fact, modeling and analyzing social networks with graph theoretic approaches, ignoring edge semantics, lead to considerable information loss. Edges may be temporal and associated to a particular event (e.g. place and time) or may hold for a particular context. Working relationships are often completely disjoint to family or friendship relationships for instance. Further, edge semantics may vary depending on the types of nodes that are connected and the type of interaction between them. The meaning of an edge linking an individual to a document could be modeled for example as "author-of" or "reviewer-of", depending on the modeler's intention. Such considerations are partially addressed by edge labeling, which however lacks semantic links to structure them.

Semantic web rich typed graph models, query languages and schema definition frameworks capture the semantics of social data. Ontologies are used to describe users and their activities, content and its relation to users. Different relationship types, trust levels and edge weights can be defined using vocabularies [1] proposing an architecture based on the Semantic Web stack, to analyze online social networks while being semantics-aware. Its purpose is to explore RDF[1]-based annotated profiles and users' interactions in social networks using background knowledge (domain vocabulary), predefined ontologies and OntoSNA, an ontology for Social Network Analysis, which provides a way to compute sociometric features with SPARQL[2].

Semantic annotation imposes structure to unstructured data, enabling better search, analysis, and information aggregation capabilities. Twitter users adopted hashtags to alleviate the significant information overload that the streaming nature of social media imposes to users interested in specific topic(s). Hashtags have been exploited for content management, organization and filtering [17, 18, 19]. Even though user-defined hashtags are ambiguous and highly heterogeneous, collaborative

[1] `http://www.w3.org/RDF/`

[2] `http://www.w3.org/TR/rdf-sparql-query/`

structures emerge [20, 21]. [18] makes use of annotated microposts together with background knowledge obtained from Linked Open Data to offer advanced search and organizational capabilities. For example, thanks to semantic links between football and sports, all information mapped only to football can be retrieved by queries about sports. Multilayered models, which involve the network between people, the network between concepts they use, and links to ontologies modeling such concepts have lately been used [22].

Figure 1 demonstrates that enterprise social network captures just one aspect of enterprise communication, representing social connectivity among employees. As employees add new friends, join groups and engage in discussions, the underlying graph structure changes frequently. Major part of business transactions are stored in enterprise databases. Such structured datasets are frequently updated with massive number of transactions and are typically used for on-line analytical processing (OLAP), business intelligence and reporting applications. Business reports, on the other hand, capture key summaries of enterprise datasets, enumerate financial information, trends, opportunities, etc. They are typically generated quarterly in the form of unstructured text, requiring some sort of automated preprocessing and analysis. Additional enterprise data sources include sensor streams, geospatial and multi-media content that exhibit varying update frequencies and formats, requiring plethora of techniques for automated processing and analysis. Regardless of the source that generates data, all datasets typically have common references of people, processes, activities, places, measures, etc. that establish linkages among them. Adding appropriate annotations in pre-processing steps facilitates integration of such heterogeneous data sources.

3 THE BLISS OF MULTIDIMENSIONALITY

Background knowledge is acquired and learned skill-set incrementally updates, interests and expertise change, as time progresses. The current focus of a specific employee may be completely different than what is stated in an outdated personal webpage or CV. We introduce temporal context (TMC) to capture such temporal effects. Further, social interactions are in many occasions bounded by, at least some, temporal and localization constraints. This refers to spatial context (SC). For instance, face-to-face interactions may only occur when individuals are physically located at the same place at the same time. Extended interactions due to office adjacency or limited communication at a conference further introduce the concept of time sensitive informal communication. Participation in meetings, talks, training, conferences, etc. constitutes event context (EC). Lengthy discussions on a daily basis indicate stronger bond than periodic hourly meetings, which in turn indicate more significance than a sporadic discussion. The relationship between two individuals therefore becomes a function of time and can be explored only as such. Static social networks ignore such interactions, establishing an edge between two users if at least some type of communication has happened, at least once. We ar-

gue that temporal correlations and causal effects between node features and social connectedness can only be manifested and magnified when considered as a function of time.

Employee interests, skills and expertise can change depending on time, work orientation and responsibilities, project focus and overall team competence. From employee's perspective, interest, expertise, curiosity, familiarity for topics constitute participation context (PPC). Topics constitute topic context (TPC). Given a context (e.g. a group discussion versus a status update) may yield significant, different aspects of employees' focus. Depending on personal or professional nature of content, different interests can be mined and different expertise levels can be identified, for disjoint set of topics. Moreover, employees often assume multiple roles in multiple projects (e.g. an employee might act as manager in one project, while being a software developer in another). This can be captured as project context (PRC). Roles and positions, and reporting hierarchy is captured in organizational context (OC). One context can be closely related to one or more other contexts. For instance, employee interests, skills and expertise can differ at multiple points in time, and be different at the same point in time within the boundaries of correspondence with different individuals.

We argue that each kind of context can be complex, thus being decomposable to "sub-contexts". In order to process enterprise communication effectively, it is imperative to establish a comprehensive model of contexts, and semantic links to structure them. Figure 2 shows various contexts in enterprise informal interactions. Various interpretations of captured scalar, hierarchical, and nominal, temporal, or spatial data that differ with context (e.g. point of view) can provide different insights or views (i.e. dimensions).

3.1 Enterprise Contextual Social Interactions

Our comprehensive list of contexts and associated dimensions enables the study and analysis of enterprise informal communication from multiple perspectives. Figure 3 depicts a scenario of enterprise contextual social interactions. Employee_AABF participates in a project (project context). In performing his/her role, s/he comes across a problem and posts a question (activity stream context) at the enterprise social networking platform, where Employee_AAAD and Employee_AADC read the question (activity stream context). Employee_AADC is interested in the problem (topic context) and starts following the question (activity stream context). Employee_AAAD responds (participation context) with a Sharepoint page reference, which was contributed by Employee_AAGG (domain context).

Figure 4 depicts information sources typically used in an enterprise. In order to find right information, employees are forced to deploy elaborate browsing strategies through a combination of multiple information mediums. The wealth of information available across mediums may become overwhelming if not properly processed, stored and presented in an intuitive way. However, mining available information sources and considering them in conjunction might lead to correlation and causality

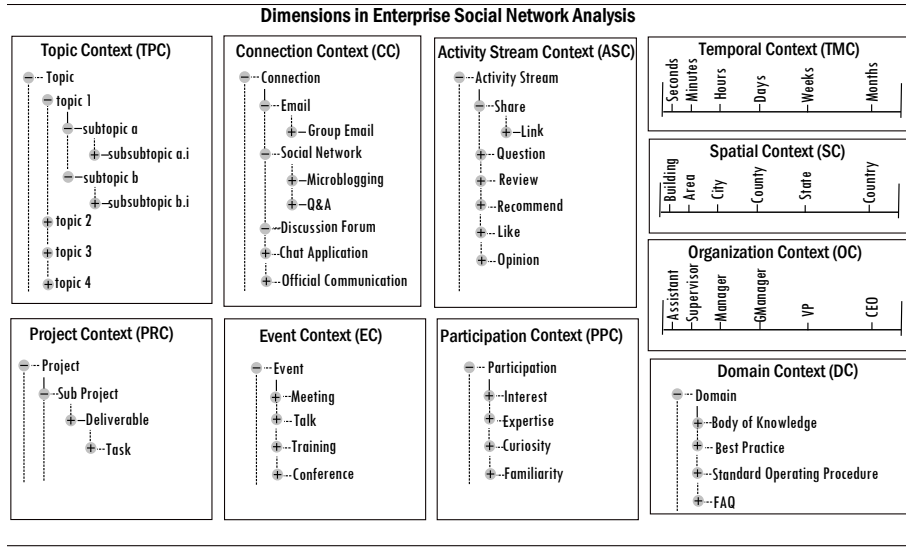

Figure 2. Contexts in enterprise social network

relations that would not otherwise become obvious. We argue that fusing available information sources is imperative for complex analysis of enterprise social communication data.

Context establishes relationships across various activities and artifacts observed in enterprise communication platforms. Once processed and annotated with appropriate contexts, activities can be retrieved as part of advanced search capability.

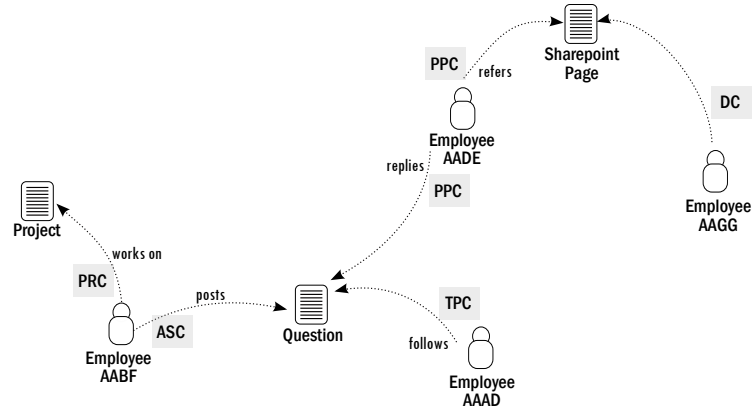

Figure 3. Contextual interactions

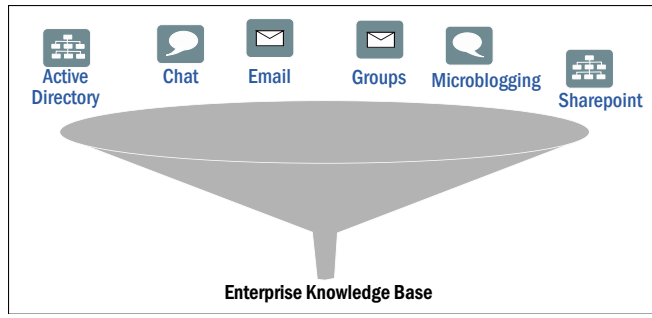

Figure 4. Information sources funneling

Figure 5 shows two such advanced search scenarios that lead to macro and micro analysis of social network content. Lens capability permits identification of users' activities in specific communication channel (e.g. Employee_AAAD's contribution in Q & A forum), thereby reveling the nature and impact of contribution. We call this capability "lens" due to its focus on a specific, narrow aspect defined by a context. However, a context can be broken down into sub-contexts in a hierarchical fashion. Topic and location, for example, can be expanded to cover subtopics and smaller location segments respectively. Query results in such case constitute a "canopy".

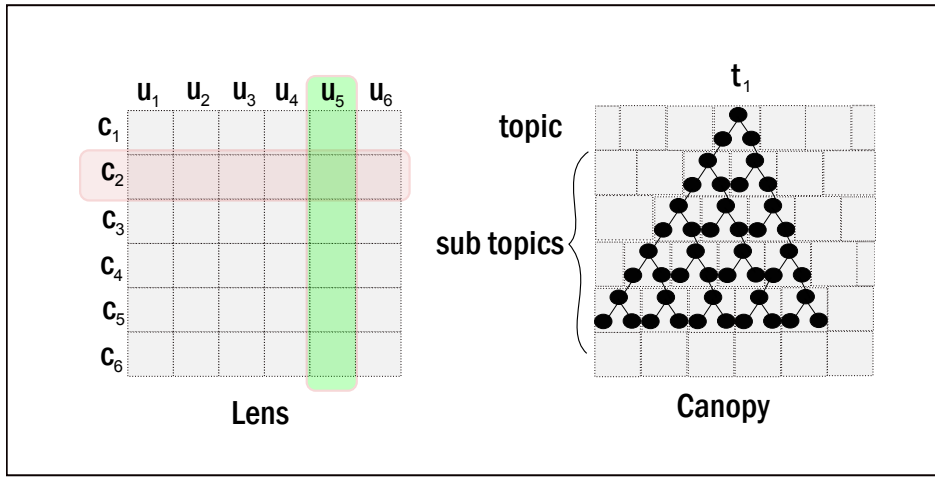

Figure 5. Lens and canopy search

4 RESONATE: SEMANTIC SOCIAL NETWORK ANALYSIS FOR AN ENTERPRISE

We propose a formal modeling that abstracts the semantics of informal communication into an integrated, context aware, time sensitive, multi-dimensional space, enabling the correlation of seemingly different domains so as to investigate them in conjunction. We introduce a novel social graph representation, shown in Figure 6, which not only contains social links between users but also maintains integrated information regarding users dynamically changing interests and activities, throughout collaboration tools used in working environments.

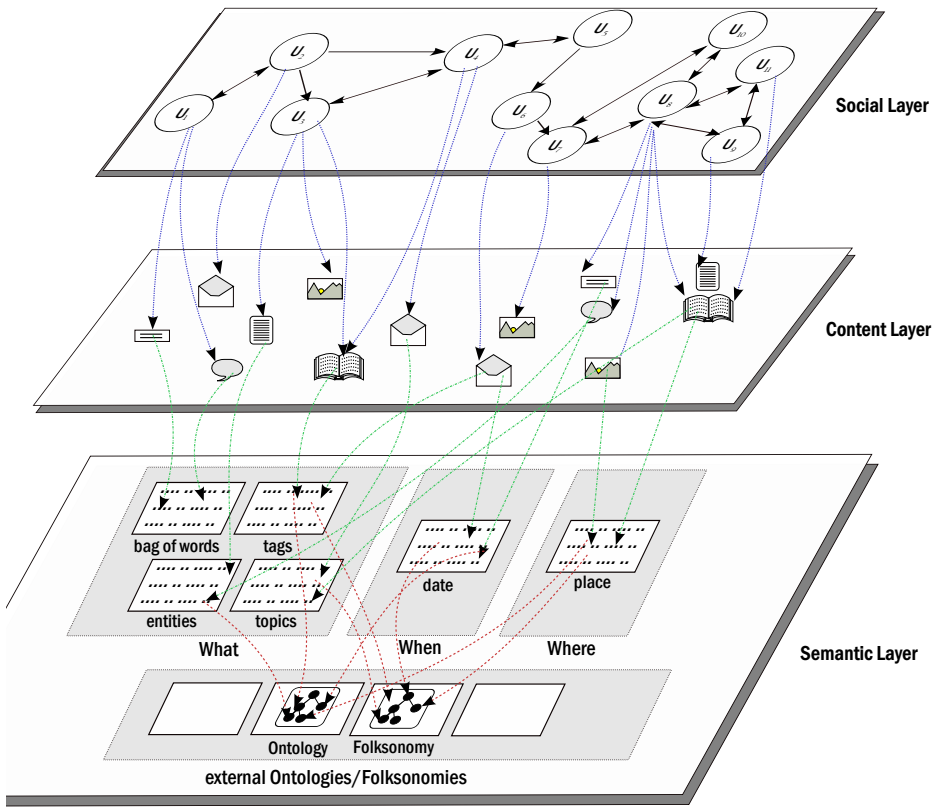

Figure 6. Layers

Social layer captures users contextual and temporal interactions. Nodes represent users and arcs represent explicit relationships (links) between them. To construct the social layer we start with friendship relationships from the social network and augment this initial graph by mining user relationships out of

available information sources: e-mail correspondence, chat networks, blogging activity, shared bookmarks and common ratings. An edge between users is defined by the context under which it is created and has an associated timestamp. A user u may be connected to user v under multiple contexts (e.g sending e-mails and social status updates) in multiple time instances.

Content layer captures published content from all available sources, including but not limited to resources shared by users (e.g. photos or videos), bookmarked and/or tagged resources (e.g. URLs), users' generated content (e.g. status updates in Facebook), e-mails, chat messages, and blog posts. Depending on available computational resources or application need, this layer may maintain raw content, which is meant to be processed later on, or in the other extreme only contain the aggregated post-processing results of previous analysis. In the latter case, provenance metadata are to be maintained in unison with analysis results so as to describe for instance the procedure followed and data sources used in the analysis.

Semantic layer contains meta-information about content, and can be broken into several constituting layers, each containing different metadata about content. This layer may include, but is not limited to, domain ontologies, vocabularies, and folksonomies and taxonomies, external sources of formal knowledge, and linked open data. OpenCalais[3], AlchemyAPI[4], and Evri[5], WordNet[6], and Freebase[7] are examples of semantic information providers and annotation enablers, exposing rich APIs for text analysis and text annotation, entity identification, and topic discovery, as well as complex relationships mining. Linked Open Data[8] can further be exploited to gain insights into knowledge that may not be inherently present in the system under examination, but is accessible through external sources.

We call our framework *Semantic Social Network Analysis for the Enterprise* (**rESONAtE**). rESONAtE enables analysis which spans layers, considering both multifaceted data and metadata, and the underlying informal communication graph. Knowledge is discovered, captured and inferred based on such complex information. In previous work [23] we define rigorous social metrics, which we use to calculate semantic similarity scores between any two object types, users and content alike, in a joint semantic space, given a context. Next, we explain how we materialize our framework and metrics to intricately model and extensively analyze a real world dataset from a Fortune 500 multinational company.

[3] `http://www.opencalais.com/`
[4] `http://www.alchemyapi.com/`
[5] `http://www.evri.com/`
[6] `http://wordnet.princeton.edu/`
[7] `http://www.freebase.com/`
[8] `http://linkeddata.org/`

We instantiate rESONAtE in the form of Ontology. The main reason we use Ontology, is that it provides a generic, reusable, and machine understandable model for representing the concepts and properties required for describing user activities and measuring their behavior. Figure 7 depicts the coverage of rESONAtE Ontology. Figure 8 shows an overview of classes in our ontology.

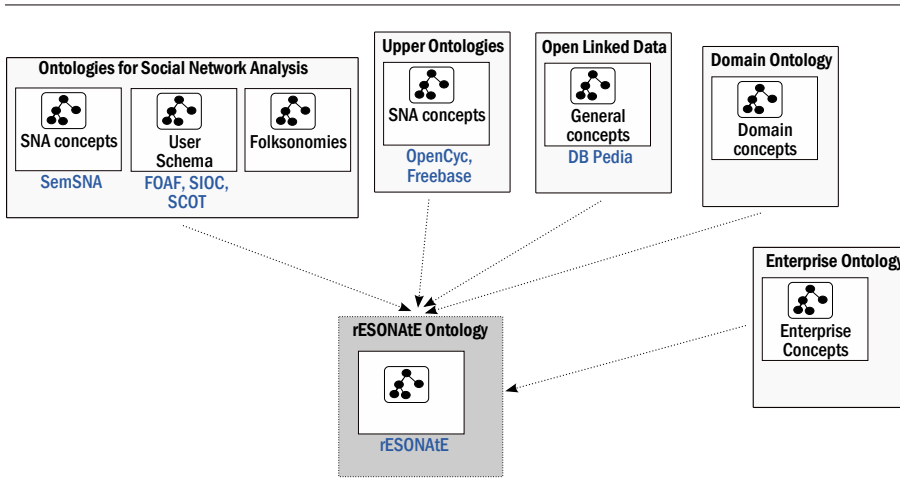

Figure 7. Enterprise social network ontology coverage

4.1 Static Modeling

`Business_Associate` class constitutes a high level abstraction of `Employee` and `Organization`. An employee may have various types of connections, in multiple contexts. For instance, an employee may have multiple supervisors while being assigned to multiple projects, and at the same time maintain a list of email contacts. An organization provides employment to workers and interacts with (e.g. sells products) other organizations. Treating `Employee` and `Organization` as direct descendants of `Business_Associate` class provides a mechanism to leverage atomic features to a collaborative level. Further, employees can interact (both explicitly and implicitly) with organizations, while organizations may participate in discussions.

`Organization` constitutes of `Companies` that contain various `Departments` and `Facilities`. Each `Employee` holds a `Position` in the enterprise and participates in one or multiple `Projects`. Unlike user profiling [24, 25], which mainly focus on static properties such as personal information, we use our ontology to leverage dynamic enterprise data. `Projects` that each employee participates in record the working status of the employee. The "StartDate" and "EndDate" describe the life span of a `Project`, based on which we can trace projects status. Moreover, it is possible for an employee to participate in multiple projects simultaneously.

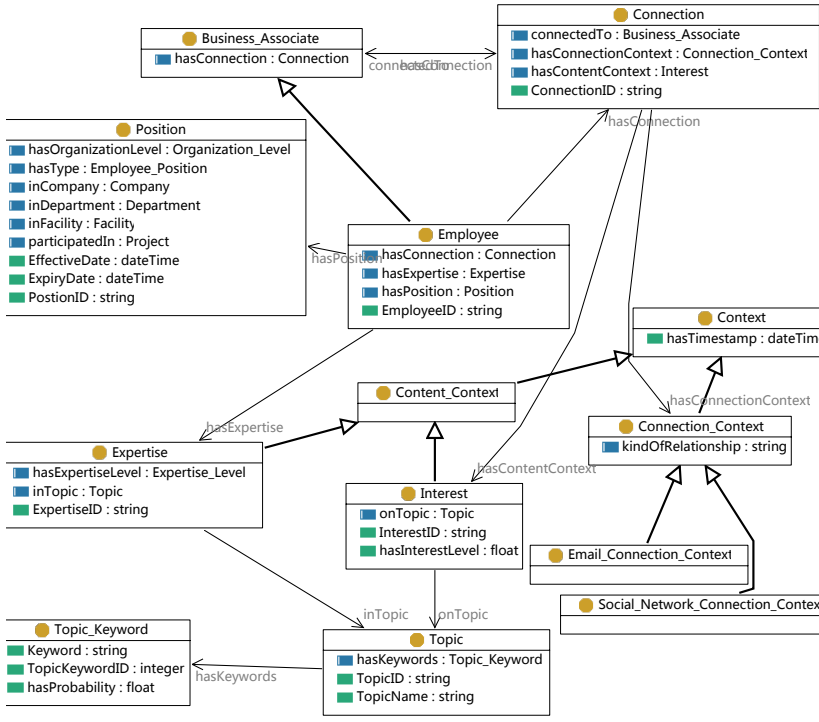

Figure 8. rESONAte ontology

`Position` class describes the current status of an employee in the enterprise and reflects on current responsibilities and focus area. It contains information such as the `Department` and `Company` for which an employee works, and in which `Facility` s/he is currently located. The exact employee position title (e.g., manager, senior engineer or CEO) is linked by "hasType" to the object class `Employee Position`. To represent the hierarchical organizational chart, we use "OrganizationLevel". "EffectiveDate" denotes the starting date of employee's position. "ExpiryDate" specifies the date that a position ends. Based on the timeline of `Positions` an employee has worked at, the complete working history of the employee can be traced back, enabling features like future position recommendation.

Accumulated skills and past experiences can be summarized from employee resumes and profiles from past appointments, even though it is difficult to objectively measure employees' expertise for different domains [26]. Previous positions and projects provide hints on employees' specialization areas, however, peer validation in the form of informal interaction acts as supporting evidence of the level of expertise of individuals. `Expertise` class is used to represent levels of expertise for each person with respect to various areas. We use the concept `Topic` (e.g., computer technology, management, accounting) to represent areas of expertise, as well

as `Interests`. Each `Topic` has a list of `Keywords` and probabilities that represent the extent to which a keyword conveys the semantic meaning of a topic.

`Connections` capture relationships between `Business_Associate` instances (i.e. `Employee` or `Organization` objects). Each connection is established under a specific context. Context can be defined as any information that can be used to characterize the situation of an entity [27]. Context can be a physical property such as time and location, or a logical concept such as a situation [28]. In our model, we consider two kinds of context: `Connection_Context` and `Content_Context`. `Connection_Context` specifies the kind of relationship a pair of entities has, such as colleagues, collaborators, or dominance/subordinate. `Content_Context` represents the common `Interest` two entities share. For example, in enterprise social media, an employee can be connected to others with follower/followee relationships. Two employees may share common interest on a `Topic` such as computer technology. Since context is a function of time, our `Context` class contains a timestamp property, which is used to record the period of time over which a context is valid.

4.2 Dynamic Modeling

We mine content of informal interactions between employees to capture their expertise in a latent topic space. Recent studies in machine learning area have developed probabilistic models to automatically uncover latent "topics" in natural language texts. Topic models [30, 32] take advantage of co-occurrence of words in text documents. They use hierarchical Bayesian models to capture the generation process of words in documents by introducing an intermediate latent topic layer. Topic models can address problems of synonymy and polysemy in natural language processing. Each topic is represented as a mixture of words with probability distribution and each document can be decomposed into a distribution of various latent topics.

In our work, we adopt the Author-Recipient Topic model (ART) [33] to discover topics in messages posted by employees in internal social media. The model builds on Latent Dirichlet Allocation [30] and the Author-Topic model [34]. Instead of only modeling topic distributions over messages or authors, the ART model conditions the distribution over topics on both the sender and the recipient of a message. Therefore, latent topics are discovered according to relationships between people. Furthermore, using ART we are able to identify not only how often employee pairs interact, but also which topics are the more prevalent ones in their discussions.

ART makes the assumption that each word w in a message is sampled from a multinomial distribution ϕ_t (the word mixture for a topic t). The topic is drawn from a multinomial distribution $\theta_i j$, which is the topic mixture specific to the author-recipient pair (i, j) of the message. We train ART using the default hyperparameters ($\alpha = 50/T$, $\beta = 200/V$), and $T = 100$ topics. Inference of ART is achieved by using 1 000-iteration Gibbs sampling [33]. We use the trained model to capture not only latent topics, but also employees' "Expertise" and "Interests" from their microblogging activity. To measure employee i's expertise on topic t, we aggregate the number of words n_{ijt} assigned to topic t and author-recipient pair (i, j), resulting

in $\sum_j n_{ijt}$. We quantify the result to a discrete scale of 1 to 5. The larger the value, the more knowledgeable an employee on a topic. Expertise is therefore a relative measure of proficiency on topic t, compared to other employees. To measure employee i's interest on topic t, we use the α smoothed normalization and obtain $\sum_j(n_{ijt}+\alpha)/\sum_t\sum_j(n_{ijt}+\alpha)$. We use this as a relative measure of the degree of preference an employee has on topic t with respect to other topics.

4.3 rESONAtE Workflow

Static and dynamic modeling of various data sources enables identification of topics, experts, connections and other relevant information across the enterprise. Expressed and latent information extracted in this manner is stored in rESONAtE. Discovered, new knowledge is provided to users in the form of recommendation of events, experts, connections, relevant content, etc. However, due to dynamism in enterprise, the modeling should be updated regularly to keep track of temporal changes, as shown in Figure 9. Ontologies derived from static modeling and SNA tasks defined in dynamic modeling are utilized to define SNA task specification in Runtime, allowing execution of SNA tasks (LDA, ART etc.) at predefined intervals.

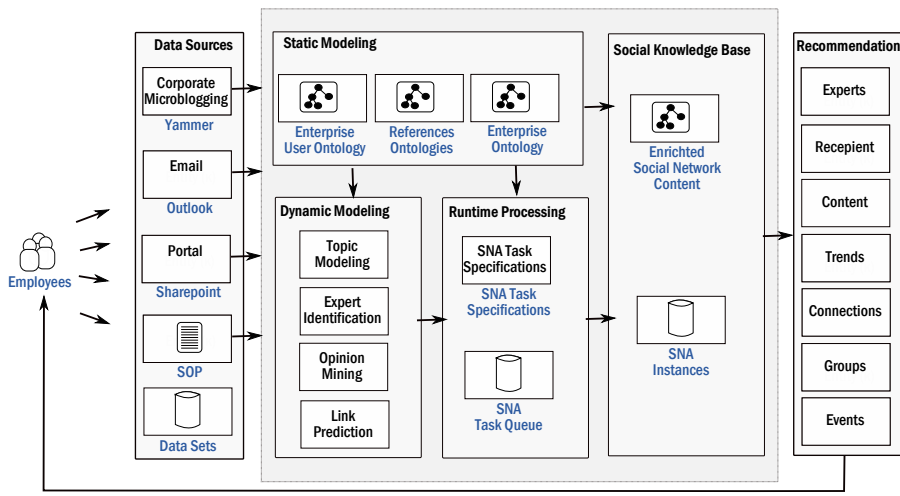

Figure 9. rESONAtE workflow

5 CASE STUDY ON REAL CORPORATE MICROBLOGGING DATA

In this section, we present a case study on a large scale dataset of corporate microblogging data from a Fortune 500 company. The functionality of the microblog-

ging service resembles that of Twitter, whereas its interface is similar to Facebook. The dataset includes 9 855 unique users, and 15 200 messages with explicit reply links to other messages, over a time span of 1.5 years. The dataset contains users activity (message exchanges) and interactions (e.g. comment/reply, tagging), but lacks explicit social relationships (followee/follower) between users. Instead, we have acquired a snapshot of the organizational hierarchy with respect to users that participate in the microblogging service. Each employee is represented with a 4-character ID (e.g. "AECF"). The organization level of each user is denoted using a character from "A" to "M". Relationships are not symmetric: employee AECF may send a message to AAAF but AAAF can choose to not reply back. Similarly, AECF may be the supervisor of AAAF. Thus we have a directed labeled graph, with multiple relations between users (i.e. multiple `Connection_Context` instances). Apart from structural closeness (i.e. being connected in the social graph), users may share common interests (see Section 4.2). We mine knowledge from message interactions between employees using the ART model (see Section 4.1).

5.1 Organizational Hierarchy vs. Informal Interactions

Organizational hierarchy is static, and dated, whereas communication may reflect "shortcuts", i.e. collaboration that spans hierarchical levels when seeking for help, or offering guidance, etc. Studying communication may reveal hidden organizational dynamics. For instance, employees Employee_AABG and Employee_AACD may belong to the same team according to organizational hierarchy, but rarely actually interact due to having diverse responsibilities.

Our intention is to better understand how information propagation works between corporate borders, and identify potential shortcuts in the organizational chart, as well as better understand how employees collaborate to tackle everyday problems and find solutions; to this end, compare employees' connections under organization hierarchy to "connections" in the microblogging service.

5.2 Contextual Ego-Network & Community Identification

Processing and analyzing employees' interactions reveals hidden dynamics. At micro level, topic oriented connections of a particular employee may reveal patterns that vary across topics. On the other hand, by clustering users with similar interests, it is possible to detect virtual communities around trending topics (macro level). A micro level query used to generate the topic gnostic ego-network of an employee follows.

5.3 Expert Identification

Employees' level of expertise may vary from topic to topic and from medium to medium. One might share innovative ideas and contribute to discussions through

Query 1: Given employee, find direct links in the organizational hierarchy.

```
PREFIX resonate:<http://local.virt/DAV/home/dav/rdf_sink/user.rdf#>
          SELECT ?Employee ?Position ?Organization_Level
 WHERE {resonate:Employee_AABF resonate:hasConection ?Connection.
             ?Connection resonate:hasconnectedTo ?Employee.
               ?Employee resonate:hasPosition ?Position.
     ?Position resonate:hasOrganizationLevel ?Organization_Level.}
```

Employee	Position	Organization_Level
Employee_AAAD	Position_2	Organization_Level_D
Employee_AAAE	Position_3	Organization_Level_E
Employee_AAAF	Position_4	Organization_Level_A

Table 1. Partial result-set of micro-analysis query listing connections of Employee_AABF, and their position in the Company

Query 2: Given employee and topic, create topic-sensitive ego-network.

```
PREFIX resonate:<http://local.virt/DAV/home/dav/rdf_sink/user.rdf#>
                       SELECT ?User ?Topic
 WHERE {resonate:Employee_AABF resonate:hasConection ?Connection.
           ?Connection resonate:hasContentContext ?Interest.
                ?Interest resonate:hasinTopic ?Topic_no.
                 ?Topic_no resonate:hasTopicName ?Topic.
              ?Connection resonate:hasconnectedTo ?User.}
```

User	Topic
Employee_AAAD	project
Employee_AAAD	connect
Employee_AAAD	pretty
Employee_AAAE	tag

Table 2. Partial result set of micro query to determine Employee_AABF's topic-oriented connections

emails, but not in microblogging sites. We differentiate expertise according to communication channels (connection context), time and content (i.e. email messages vs. microposts). A micro level query that retrieves topics and levels of expertise for given employee follows. Expertise levels are quantized, taking integer values between 1 (less expert) and 5 (authority).

5.4 Trends Macro-Analysis

Discovery of trending topics is a typical application in social network analytics. Typically, trending topics are identified for a single communication channel for single

Query 3: Given employee, retrieve her areas of expertise and expertise levels.

```
PREFIX resonate:<http://local.virt/DAV/home/dav/rdf_sink/user.rdf#>
                    SELECT ?Topic ?Name ?Level
  WHERE {resonate:Employee_AABG resonate:hasExpertise ?Expertise.
                 ?Expertise resonate:hasinTopic ?Topic.
                   ?Topic resonate:hasTopicName ?Name.
              ?Expertise resonate:hasExpertiseLevel ?Level.}
```

Topic	**Name**	**Level**
Topic_97	share	Expertise_Level_1
Topic_84	drive	Expertise_Level_1
Topic_9	test	Expertise_Level_1

Table 3. Partial result set of Employee_AABG's expertise areas and levels

network. Such analysis at macro level may not be sufficient for an enterprise. In enterprise context, such macro analysis can prove to be more useful by discovering, for instance, trending topics and trending users for each communication channel, department, and organizational position. Given a trending topic, the following query retrieves its most prominent keywords, in all contexts. This is direct result of performing dynamic modeling over the enriched corpus.

Query 4: Given topic, get probability distribution of its trending keywords.

```
PREFIX resonate:<http://local.virt/DAV/home/dav/rdf_sink/user.rdf#>
                    SELECT ?Keyword ?Probability
  WHERE {resonate:Topic_97 resonate:hasKeyword ?Topic_Keyword_No.
              ?Topic_Keyword_No resonate:hasKeyword ?Keyword.
        ?Topic_Keyword_No resonate:hasProbability ?Probability.}
```

Keyword	**Probability**
share	0.24534
interest	0.15791
friend	0.04047
facebook	0.0396
story	0.02046
easy	0.01176
found	0.01002
quick	0.00915
intern	0.00872
earlier	0.00741

Table 4. Keywords and probability distribution for given topic

6 CONCLUSION

In this paper we argued that enterprise dynamism leads to big data challenges. We proposed to address integration and knowledge management aspects with a combination of semantic web techniques and social network analytics, capable of handling latent and expressed semantics in the enterprise. We argued that in a typical enterprise, knowledge is always expressed and utilized in some context. We identified various kinds of contexts and proposed a multidimensional model, which covers both static and dynamic communication aspects. We proposed a knowledge management workflow that enables, among others, continuous discovery of trends, expertise and interests, both at the employee and the enterprise level. Finally, we discussed illustrative use cases that demonstrate how our approach can be useful in practical enterprise setting. Our initial experiments on a representative microblogging dataset indicated vast potential of semantic social network analysis in addressing big, multidimensional data challenges for the enterprise. Our future work will focus on

1. integrating and experimenting with more communication channels,
2. deploying and extending social network analytic techniques, and
3. building advanced visualization capabilities.

REFERENCES

[1] Ereteo, G.—Limpens, F.—Gandon, L.—Corby, F. O.—Buffa, M.—Leitzelman, M.—Sander, P.: Semantic Social Network Analysis: A Concrete Case. In Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena, IGI Global, January 2011, pp. 122–156.

[2] Cross, R.—Parker, A.—Borgatti, S. P.: A Bird's-Eye View: Using Social Network Analysis to Improve Knowledge Creation and Sharing. Knowledge Directions, Vol. 2, 2000, No. 1, pp. 48–61.

[3] Robinson, M. A.: Erratum: Correction to Robinson, M. A.: An Empirical Analysis of Engineers' Information Behaviors. Journal of the American Society for Information Science and Technology, Vol. 61, 2010, pp. 1947–1947.

[4] Patri, O. P.—Sorathia, V.—Prasanna, V. K.: Event-Driven Information Integration for the Digital Oilfield. In SPE Annual Technical Conference and Exhibition (ATCE), Society of Petroleum Engineers (SPE), October 2012.

[5] Tuulos, V. H.—Perkiő, J.—Tirri, H.: Multi-Faceted Information Retrieval System for Large Scale Email Archives. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05), New York, NY, USA, ACM 2005, pp. 683–683.

[6] Burkhart, T.—Werth, D.—Loos, P.: Context-Sensitive Business Process Support Based on Emails. In Proceedings of the 21st International Conference Companion on World Wide Web (WWW '12 Companion), New York, NY, USA, ACM 2012, pp. 851–856.

[7] Gloor, P. A.—Laubacher, R.—Dynes, S. B. C.—Zhao, Y.: Visualization of Communication Patterns in Collaborative Innovation Networks – Analysis of Some W3C Working Groups. In Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM '03), New York, NY, USA, ACM, 2003, pp. 56–60.

[8] Diesner, J.—Frantz, T.—Carley, K.: Communication Networks from the Enron Email Corpus It's Always About the People. Enron Is No Different? Computational & Mathematical Organization Theory, Vol. 11, 2005, pp. 201–228.

[9] Gentile, A. L.—Lanfranchi, V.—Mazumdar, S.—Ciravegna, F.: Extracting Semantic User Networks from Informal Communication Exchanges. In Proceedings of the 10th International Conference on the Semantic Web 2011, Part I, Springer-Verlag 2011, pp. 209–224.

[10] Wu, A.—DiMicco, J. M.—Millen, D. R.: Detecting Professional Versus Personal Closeness Using an Enterprise Social Network Site. In Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI '10), New York, NY, USA 2010, pp. 1955–1964.

[11] Wang, Q.—Jin, H.—Liu, Y.: Collaboration Analytics: Mining Work Patterns from Collaboration Activities. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10), New York, NY, USA, 2010, pp. 1861–1864.

[12] Chelmis, C.—Prasanna, V. K.: Social Networking Analysis: A State-of-the-Art and the Effect of Semantics. In The Third IEEE International Conference on Social Computing (SocialCom 2011), October 2011, pp. 531–536.

[13] Scott, J.: Social Network Analysis: A Handbook. Sage 2000.

[14] Wasserman, S.—Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press 1994.

[15] Mika, P.: Ontologies Are Us: A Unified Model of Social Networks and Semantics. In International Semantic Web Conference 2005, pp. 522–536.

[16] Au Yeung, C. M.—Gibbins, N.—Shadbolt, N.: Mutual Contextualization in Tripartite Graphs of Folksonomies. In The 6th International Semantic Web Conference (ISWC '07), November 2007, pp. 966–970.

[17] Amer-Yahia, S.—Huang, J.—Yu, C.: Building Community-Centric Information Exploration Applications on Social Content Sites. In Proceedings of the 35th SIGMOD International Conference on Management of Data (SIGMOD '09), pp. 947–952.

[18] Mendes, P. N.—Passan, A.—Kapanipathi, P.—Shet, A. P.: Linked Open Social Signals. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT '10, Vol. 01, IEEE Computer Society 2010, pp. 224–231.

[19] Huang, J.—Thornton, K. M.—Efthimiadis, E. N.: Conversational Tagging in Twitter. In Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT '10), ACM 2010, pp. 173–178.

[20] Wagner, C.: Exploring the Wisdom of the Tweets: Towards Knowledge Acquisition from Social Awareness Streams. In The Semantic Web: Research and Applica-

tions, Lecture Notes in Computer Science, Springer 2010, Vol. 6089, pp. 493–497, 10.1007/978-3-642-13489-0_50.

[21] Trant, J.: Studying Social Tagging and Folksonomy: A Review and Framework. Journal of Digital Information, Vol. 10, 2009, No. 1.

[22] Jung, J. J.—Euzenat, J.: Towards Semantic Social Networks. In Proceedings of the 4th European Conference on the Semantic Web: Research and Applications (ESWC '07), Springer-Verlag 2007, pp. 267–280.

[23] Chelmis, C.—Sorathia, V.—Prasanna, V. K.: Enterprise Wisdom Captured Socially. In The 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Istanbul, Turkey 2012.

[24] Golemati, M.—Katifori, A.—Vassilakis, C.—Lepouras, G.—Halatsis, C.: Creating an Ontology for the User Profile: Method and Applications. In Proceedings of the First RCIS Conference 2007, pp. 407–412.

[25] Tazari, M.—Grimm, M.—Finke, M.: Modelling User Context. In Proceedings of 10th International Conference on Human-Computer Interaction 2003.

[26] Fink, J.—Kobsa, A.—Nill, A.: Adaptable and Adaptive Information Access for all Users, Including the Disabled and the Elderly. In International Conference UM97, Wien New York, Springer, Vol. 171, 1997, p. 173.

[27] Abowd, G.—Dey, A.—Brown, P.—Davies, N.—Smith, M.—Steggles, P.: Towards a Better Understanding of Context and Context-Awareness. In Handheld and Ubiquitous Computing, Springer 1999, pp. 304–307.

[28] Stan, J.—Egyed-Zsigmond, E.—Joly, A.—Maret, P. et al.: A User Profile Ontology for Situation-Aware Social Networking. In Proceedings of 3rd Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI2008), 2008.

[29] Jung, J. J.: Service Chain-Based Business Alliance Formation in Service-Oriented Architecture. Expert Systems with Applications, Vol. 38, 2011, No. 3, pp. 2206–2211.

[30] Blei, D.—Ng, A.—Jordan, M.: Latent Dirichlet Allocation. The Journal of Machine Learning Research, Vol. 3, 2003, pp. 993–1022.

[31] Jung, J. J.: Evolutionary Approach for Semantic-Based Query Sampling in Large-Scale Information Sources. Information Sciences, Vol. 182, 2012, No. 1, pp. 30–39.

[32] Steyvers, M.—Griffiths, T.: Probabilistic Topic Models. Handbook of Latent Semantic Analysis, Vol. 427, 2007, No. 7, pp. 424–440.

[33] McCallum, A.—Wang, X.—Corrada-Emmanuel, A.: Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. Journal of Artificial Intelligence Research, Vol. 30, 2007, No. 1, pp. 249–272.

[34] Rosen-Zvi, M.—Griffiths, T.—Steyvers, M.—Smyth, P.: The Author-Topic Model for Authors and Documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, AUAI Press 2004, pp. 487–494.

Charalampos Chelmis is a Ph. D. candidate in the Department of Computer Science at University of Southern California. He received his Master of Science in Computer Science degree from the Computer Science, University of Southern California in 2010 and his Diploma in Computer Engineering and Informatics from CEID, University of Patras, Greece in 2007. His research interests include complex social networking analysis, social networking data mining and analysis, semantically enriched social network modeling and analysis, collective intelligence, social computing, information networks, big data analytics, enterprise microblogging, semantic web and its applications to energy informatics.

Hao Wu is currently a Ph. D. student in computer science at University of Southern California. He received BS and MS degrees from Zhejiang University, China. His research topic is on application of machine learning to social media analytics.

Vikram Sorathia is a postdoctoral research associate in the Ming Hsieh Department of Electrical Engineering at the University of Southern California. He received his Ph. D. degree from the Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), India in 2009. His research interests include services science, knowledge management, big data analytics and software architecture frameworks.

Viktor K. Prasanna is Charles Lee Powell Chair in Engineering in the Ming Hsieh Department of Electrical Engineering and Professor of Computer Science at the University of Southern California. His research interests include high performance computing, parallel and distributed systems, reconfigurable computing, and embedded systems. He received his B. Sc. in Electronics Engineering from the Bangalore University, M. Sc. from the School of Automation, Indian Institute of Science and Ph. D. in Computer Science from the Pennsylvania State University. He is the Executive Director of the USC Infosys Center for Advanced Software Technologies (CAST) and is an Associate Director of the USC-Chevron Center of Excellence for Research and Academic Training on Interactive Smart Oilfield Technologies (Cisoft). He also serves as the Director of the Center for Energy Informatics at USC. He

served as the Editor-in-Chief of the IEEE Transactions on Computers during 2003–2006. Currently, he is the Editor-in-Chief of the Journal of Parallel and Distributed Computing. He was the Founding Chair of the IEEE Computer Society Technical Committee on Parallel Processing. He is the Steering Co-Chair of the IEEE International Parallel and Distributed Processing Symposium (IPDPS) and is the Steering Chair of the IEEE International Conference on High Performance Computing (HiPC). He is a fellow of the IEEE, the ACM and the American Association for Advancement of Science (AAAS). He is a recipient of the 2009 Outstanding Engineering Alumnus Award from the Pennsylvania State University.