Semantic Self-Formation of Communities of Peers [*]

Silvana Castano and Stefano Montanelli

Università degli Studi di Milano
DICo - Via Comelico, 39, 20135 Milano - Italy
{castano,montanelli}@dico.unimi.it

Abstract. The formation of semantic communities of peers plays a crucial role for realizing effective query propagation mechanisms on a semantic basis. In this paper, we propose a novel approach to the self-organization of autonomous communities of peers; we propose *semantic handshake techniques* based on semantic community aggregation and *community-aware query propagation techniques* exploiting dynamic ontology matching techniques for improving traditional P2P search and discovery capabilities.

1 Introduction

Schema-based P2P networks [8, 9] go beyond traditional file-sharing P2P networks, by providing infrastructures where peers can share and create knowledge. In such systems, peers act as autonomous and independent agents and share knowledge by submitting discovery queries and by replying with relevant knowledge. The self-formation and management of semantic communities [10] and the availability of advanced techniques for query propagation on a semantic basis is a challenging issue in the current stage of development of open networked system architectures and schema-based P2P networks, to enforce sharing of distributed resources and semantic collaboration in an effective way. To this end, *ontologies* are generally employed for describing the knowledge to be shared, and appropriate techniques for *consensus negotiation* are required to deal with the different concept meanings in the ontologies provided by different peers for community formation. In this paper, we address the problem of formation of semantic communities of peers and we describe the work we have recently undertaken in the framework of our schema-based HELIOS P2P network [1]. In particular, after presenting the ontology matching foundations for supporting semantic communities in P2P systems (Section 2 and 3), we propose a novel approach to the self-organization of autonomous communities of peers based on i) *Semantic handshake techniques* to handle the problem of consensus

[*] This paper has been partially funded by "Wide-scalE, Broadband, MIddleware for Network Distributed Services (WEB-MINDS)" FIRB Project funded by the Italian Ministry of Education, University, and Research, and by NoE INTEROP, IST Project n. 508011 - 6th EU Framework Programme.

negotiation and agreement to commit a declaration of interest to a semantic community, by matching ontologies [6, 7] for organizing the committed peers according to a structured organization for efficient query propagation (Section 4); ii) *Community-aware query propagation techniques* for supporting effective peer communications at intra/inter community level and for improving traditional P2P search and discovery P2P capabilities (Section 5). Finally, we will compare the presented techniques with main related work in literature (Section 6) and we will discuss our future research work (Section 7).

2 Foundations of semantic communities

In a P2P system, each peer acts as an autonomous and independent agent and shares knowledge by submitting discovery queries and by replying with relevant knowledge. In this context, the role of semantic communities of peers is related to the capability of dynamically aggregating nodes with similar interests in efficient structures in order to i) reduce the network load due to overlapping single peer requests and ii) define effective communication mechanisms for sets of nodes which share the same understanding of a domain of interest (i.e., peers members of the same community).

We define a *semantic community of peers* as a set of nodes which show a common interest in a given topic and are organized in a structured way (e.g., a tree). Semantic communities are *autonomously emerging*, in that they originate from a declaration of interest of a peer and group those peers which spontaneously agree with the declaration, since they have relevant resources for the community. A community of peers is identified by a unique *Community Identifier* (*CID*), and a subject category or topic area of interest called community *Identity Card* (*ICard*), defined as an ontology. The use of an ontology-based ICard provides a semantically rich description of a given topic area of interest and allows the characterization of the common interpretation (i.e., perspective) of the topic area featuring the community. The following foundations characterize the formation of semantic communities:

- *Ontology-based peer description.* Each peer exposes to the system a peer ontology which provides a semantically rich representation of the resources that the peer exposes to the network, in terms of concepts, properties, and semantic relations.
- *Query-based interactions.* Each peer interacts in a peer-to-peer manner with the other members of the system by submitting discovery queries in order to identify the potential members of a given community and by replying to incoming queries whether it can join a given community.
- *Semantic matchmaking capabilities.* Each peer implements a *semantic matchmaker* for matching ontologies in order to find which concepts match in different ontologies and at which level.

We address the semantic community formation process under the constraints that: i) each peer can be member of multiple communities and stores the CID and the ICard of each joined community ii) no centralized authority (e.g., Super-Peer) is expected to coordinate the community discovery and formation process, and iii) the choice of joining an emergent community with a given ICard depends on the semantic matchmaker results. Receiving an incoming community ICard (i.e., an ontology), a peer invokes the semantic matchmaker and compares such an ICard with its peer ontology. A peer joining the community C_1 is required to provide concepts in its peer ontology with a high semantic affinity with the ICard of C_1.

In the following, we describe semantic community formation techniques based on the semantic matchmaker and related matching techniques we have developed in the framework of Helios (Helios EvoLving Interaction-based Ontology knowledge Sharing) P2P systems [1].

3 Ontology matching in Helios

Helios is a system for ontology-based knowledge discovery and sharing in peer-based open distributed systems. Helios is a multi-ontology environment where each peer provides its own peer ontology and uses a semantic matchmaker in order to identify the semantic affinity among concepts stored in the peer ontologies of different peers. Each peer ontology is stored in a metadata repository organized according to H-Model [7], a language independent ontology model capable of describing a number of ontology specification formalisms (e.g., OWL, RDF(S), UML) in a Semantic Web-compatible manner, in terms of concepts, properties, and semantic relations. In Helios, the semantic matchmaker is based on the H-Match algorithm [7] performing dynamic ontology matching by taking into account both linguistic and contextual features of the concepts to be compared. H-Match performs ontology matching at different levels of depth, implementing four different *matching models* spanning from surface to intensive matching, with the goal of providing a wide spectrum of metrics suited for dealing with many different matching scenarios that can be encountered in comparing concept descriptions of real ontologies. The *surface matching* is defined to consider only the names of concepts. Surface matching is suited for dealing with high-level, poorly structured ontological descriptions. The *shallow matching* is defined to consider both concept names and concept properties. With this model, we want a more accurate level of matching, by taking into account not only the concept names but also information about the presence of properties and about their cardinality constraints. The *deep matching* model is defined to consider concept names and the whole context of concepts, by considering also semantic relations. Finally, the *intensive matching* model is defined to consider, in addition to the features of the deep model, also property values, for providing the highest accuracy in semantic affinity evaluation. Each model calculates a *semantic affinity value* $SA(c, c')$ of two concepts c and c' which expresses their level of matching by considering linguistic and contextual features of concept descriptions. $SA(c, c')$ is based on a linguistic affinity coefficient $LA(c, c')$ and on a contextual affinity coefficient $CA(c, c')$. The linguistic affinity coefficient $LA(c, c') \in [0, 1]$ between

two concepts c and c' calculates their level of matching based on the meaning of their names n_c and $n_{c'}$. To this end, the lexical system WordNet is exploited for defining a thesaurus Th of terms and terminological relationships between names. For Th construction, a subset of the WordNet terminological relationships is considered (i.e., synonymy, hypernymy, hyponymy, meronymy, holonymy, coordinated term). $LA(c, c')$ is computed by assigning a weight to each considered terminological relationship and by calculating the highest-strength path of terminological relationships between n_c and $n_{c'}$ if at least one path exists, otherwise $LA(c, c')$ is zero. Path strength is computed by multiplying the weights associated with each terminological relationship involved in the path.
The contextual affinity coefficient $CA(c, c') \in [0, 1]$ between two concepts c and c' intends to capture their affinity based on their the contexts Ctx_c and $Ctx_{c'}$. Depending on the matching model, Ctx_c and $Ctx_{c'}$ can be composed only by properties (shallow) or by properties and semantic relations (deep). Furthermore, when the intensive matching is adopted, property values are also considered in the computation of the contextual affinity coefficient. $CA(c, c')$ is proportional to the number of matching elements the concepts have in their contexts and to their level of matching. The level of matching of two context elements is computed by considering their linguistic affinity and the kind of relation they have with c and c', respectively. In the computation of $CA(c, c')$, each context element $e \in Ctx_c$ is compared with each context element $e' \in Ctxc'$ in order to evaluate their level of matching. For a given element e, if only one matching element e' is identified, the corresponding matching value between e and e' is considered for the computation of CA. In case that more than one matching element e' is identified, the best matching pair (e, e') is considered for the evaluation of $CA(c, c')$, that is, the pair with the highest matching value. If no matching element e' is found, the best matching value for e is set to zero (i.e., e is not considered for CA computation). Finally, CA is computed as the ratio of the sum of the best matching value for each $e \in Ctx_c$ to the number of elements considered in the context of c (i.e., the cardinality of Ctx_c).
Finally, the semantic affinity $SA(c, c') \in [0, 1]$ is evaluated as follows:

$$SA(c, c') = W_{LA} \cdot LA(c, c') + (1 - W_{LA}) \cdot CA(c, c') \tag{1}$$

where the relevance of the linguistic and the contextual features in the semantic affinity evaluation process is established, by setting the weight $W_{LA} \in [0, 1]$. A detailed description of H-Match and related matching models can be found in [6, 7].

4 Self-formation of semantic communities of peers

In this section, we present a *semantic handshake* process based on consensus negotiation techniques for the formation of semantic communities of peers. Moreover, we discuss a running example in order to stress the role of H-Match and related ontology matching techniques in the consensus negotiation process.

4.1 The community formation process

A semantic community of peers emerges when a node, called *community founder*, invokes a *semantic handshake* process which is composed of the following tasks: *ICard advertisement*, *member identification*, *request approval*, and *community commitment*.

ICard advertisement. The founder P_f defines a CID and an ICard describing the topic area of interest of the emerging community, along with a set of *commitment constraints* specifying the conditions required for the community establishment (e.g., minimum number of member required, specific semantic affinity constraints). Then, the founder composes an *Invitation Message* containing the CID and the ICard created, as well as the TTL parameter defining the maximum number of hops allowed for the invitation propagation, the matching model to be used for affinity evaluation (i.e., surface, shallow, deep, intensive), and the matching threshold t specifying the minimum semantic affinity value required to consider a concept of the ICard and a concept of a peer ontology as matching concepts. Then, the invitation message is sent to all P_f neighbors in order to advertise the new community.

Member identification. Each invited peer P_i invokes the semantic matchmaker in order to compare the incoming ICard with its peer ontology. P_i is relevant for the community if the semantic matchmaker identifies concepts in the peer ontology with a semantic affinity higher than the specified threshold t with respect to the ICard. In this case P_i replies to P_f with an *Interest Message* reporting the portion of its peer ontology related to the matching concepts found to be relevant for the community by the semantic matchmaker. Independently from the matchmaker results and if $TTL \geq 0$, P_i forwards the invitation message to all its neighbors, except for the peer from which the message has been received. Each invited peer discards duplicate copies of the same invitation message possibly received.

Request approval. Receiving the interest messages, the founder P_f has to evaluate which peers are admitted in the community. For this reason, P_f invokes its semantic matchmaker and compares each peer ontology portion received by the interested peers with its knowledge (i.e., its peer ontology). For each candidate peer, the goal of this comparison is to evaluate whether the provided knowledge matches the knowledge of the founder, and then to assess whether they share a common perspective of the community interests. If the matchmaker returns matching results higher than t, P_f admits the peer in the community and sends an *Approval Message* to the admitted peer.

Community commitment. Once the Request approval phase is completed, the founder verifies that the commitment constraints are satisfied. In this case, a *Commitment Message* is sent to all the admitted peers and the semantic community is effectively established. If the committed constraints are not satisfied, the founder stops the community formation. In this case, the admitted peers wait for the commitment message until a predefined timeout expires and the community is considered as disbanded.

As an example, consider Figure 1 where the handshake algorithm is applied to a snatch of a P2P network and the community founder, represented by a double hoop, sets an initial $TTL = 2$. In Figure 1, dashed lines represent random P2P connections and the path followed by the invitation message (continuous line) defines a tree structure where the root is identified by the community founder and the leafs are represented by the invited peer with $TTL = 0$. Each invited peer negotiates its participation in the community directly with the community founder. Once it is admitted, the peer exploits the tree structure and communicates within the community through its community neighbors. We define the *community neighbors* of a community member P_m as the peer that invited P_m in the community (i.e., P_m predecessor) and the peers that P_m invited in the community (i.e., P_m successors). An invited peer not interested in the community or discarded by the founder is to be pruned from the tree structure of the community. For this reason, after the approval phase, each community member P_m notifies to its predecessor P_p of its presence in the community. If P_p is not member of the community, it forwards the P_m notification to its predecessor P_g and notifies P_m that P_g is its new predecessor.
As an example, consider peer E in Figure 1. The community members peer H and peer K notify peer E of their participation. Peer E has not joined the community and is to be pruned from the community tree. Then, peer E forwards the notification to peer B and notifies peer H and peer K that peer B is their new predecessor.

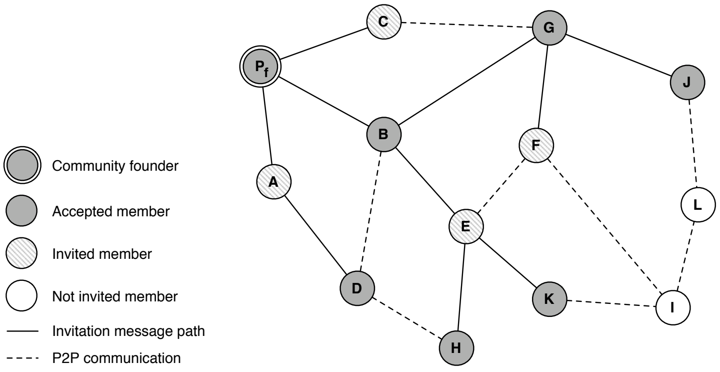

Fig. 1. Example of aggregation of a semantic community

In Figure 2, the state transition diagram for the handshake process is described. We observe that the aggregation of a semantic community of peers passes through the following states. At the beginning, the semantic community is expressed at a potential level and lies in the *potential community* state. When the community founder defines the ICard and CID, the community starts the ICard advertisement transition and moves in the *emerging community* state in which the invited peers are called to show their interest in the rising community. The community remains in an emerging state until the invitation message is propagated to all the invited peers and the identification member transition is completed. With the request approval transition, the community moves in the *partially committed* state where the accepted peers are notified of their membership. Only after the completion of the commitment transition, the semantic community enters the *committed community* state and becomes effective in the network.

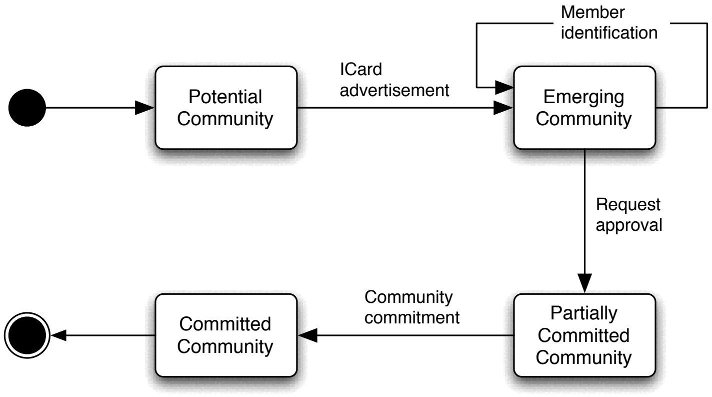

Fig. 2. The state transition diagram of the handshake algorithm

4.2 Running example

In order to put in evidence the role of ontology matching in the community aggregation process, we consider the example of Figure 3 where we show a portion of the network of Figure 1 and we discuss the role of the H-MATCH algorithm during the ICard advertisement and the member identification phases. In Figure 3, peers are represented together with a portion of their peer ontologies described with H-MODEL. The peer ontology of peer P_f is related to the publishing domain. Let us assume that the Peer P_f is interested in defining a semantic community with an ICard containing the concept Publication with the properties author and title. Then, P_f composes an invitation message with such an ICard together with a $TTL = 1$, the deep matching model to be used, and a matching threshold $t = 0.5$. Such an invitation message is sent to the topological neighbors of peer P_f (i.e., peer A, peer B, peer C) and, according to the

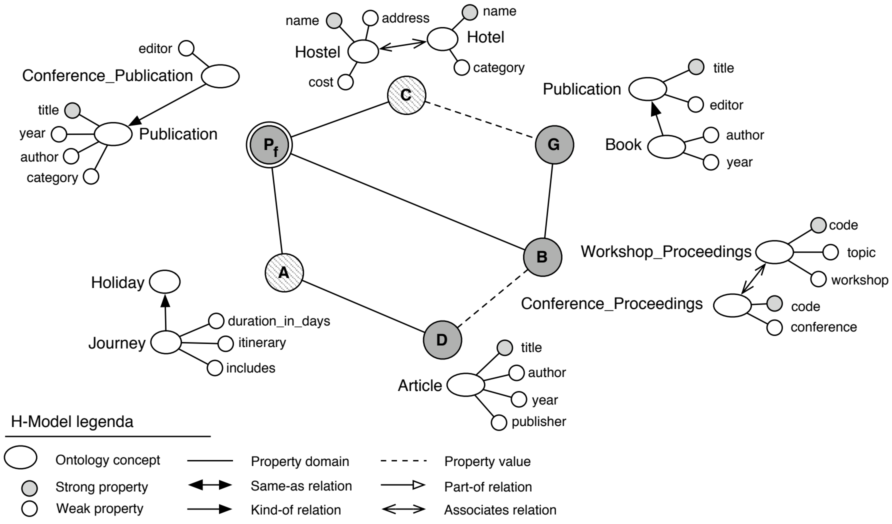

Fig. 3. Example of network with peers and associated peer ontologies

TTL, the invitation is also forwarded to peer D and peer G. Each receiving peer invokes the H-Match algorithm with the deep matching model to evaluate the semantic affinity between the incoming ICard and its respective peer ontology. The peer ontologies of peer A and peer C are related to the tourism domain and no matching concepts are found for the ICard. For this reason, these peers do not reply to the invitation message and are no more considered for the community aggregation. For what concern peer B, peer D, peer G, the H-Match results produced with the deep model are reported in Table 1 [1]. According the

Peer	H-Match result
peer B	$SA(Publication, Conference\_Proceedings) = 0.56$ $SA(Publication, Workshop\_Proceedings) = 0.56$
peer D	$SA(Publication, Article) = 0.86$
peer G	$SA(Publication, Book) = 0.71$ $SA(Publication, Publication) = 0.85$

Table 1. The H-Match results for peer B, peer D, peer G

[1] We note that comparing the concept Publication in the ICard and the concept Publication in peer G ontology, H-Match produces $SA(Publication, Publication) = 0.85$. This is due to the fact that the two concepts Publication have different contexts and, this affects the computation of CA coefficient when the deep model is adopted.

threshold $t = 0.5$ specified in the invitation, peer B, peer D, peer G can provide relevant concepts for the ICard, and reply to peer P_f with an interest message. Then, peer B, peer D, peer G will be considered by peer P_f during the request approval and community commitment phases for the definition of the committed community.

5 Community-aware query propagation

Committed communities are the reference for improving search and discovery capabilities in P2P networks. When a searching peer P_s submits a discovery query to the system for discovering relevant resources matching the target specified in the query, recipients have to be identified to avoid the flooding of the request. The communities of peers are exploited for query dissemination by addressing each request to the set of recipients that can provide resources matching the target. To this end, P_s exploits its joined communities in order to discover whether their ICards are related to the query target. P_s invokes the semantic matchmaker and evaluates the semantic affinity between an incoming query Q and the ICard of each joined community. On the basis of H-Match results, we distinguish the following cases:

- P_s is member of one or more communities related to the query Q. For each community found to be relevant, P_s sends the query Q to its semantic neighbors in the community. Each receiving node P_r forwards the query Q to its community neighbors except for P_s, and invokes its semantic matchmaker to compare the query Q against its peer ontology in order to evaluate whether it can provide relevant knowledge to send back to P_s. The forwarding mechanism is iterated until the query Q reaches each community member.
- No semantic affinity exists between the query Q and the ICard of the communities to which P_s belongs. Q is sent to all the peers known by P_s according to the routing protocols of the underlying P2P infrastructure. Each receiving peer invokes its semantic matchmaker and compares the contents of Q with the ICard they own in order to renew the community-aware query propagation.

As an example of community-aware query propagation, we consider the semantic community defined in the previous example of Figure 1. In Figure 4, we show the tree structure of such a semantic community where we assume that the peer K, on the basis of its semantic matchmaker results, needs to submit a query to this community. As shown in Figure 1, peer K sends the query to peer B (instant 1) and peer B forwards the query to its community neighbors (instant 2), namely peer P_f, peer G, and peer H. Finally, the query is forwarded to peer D and peer J (instant 3) by peer P_f and peer G, respectively.

6 Related Work

In P2P systems, the role of semantic communities for improving search and discovery techniques is crucial due to the dynamism of peers and their unpre-

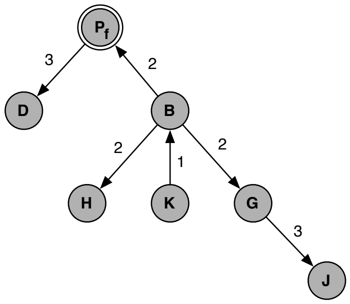

Fig. 4. Example of community-aware query propagation

dictable requirements and to the lacks in semantically rich representation of knowledge to be shared. In literature, some relevant works have been appeared with respect to semantic community management in P2P systems. For instance, in [3] the authors present a P2P architecture for supporting peer federations in which knowledge sharing is based on a social collaboration model. Single peers and federations use the Kex platform (Knowledge Exchange System) for organizing knowledge from an individual- or community-based perspective managing different meanings by means of a semantic matching algorithm. The formation of dynamic coalitions for pruning the search space and allow a better dissemination of information to participating peers is discussed in [12]. In this approach, communities emerge autonomously according to the interests advertised by peers and each peer is characterized by an involvement attribute which state the level of participation of the peer in the community.

Semantic communities of peers can have an impact on the performance of P2P discovery and search processes. The problem of a semantic query propagation in P2P systems is being considered and some projects are being developing semantic-oriented query routing approaches based on metadata and on a notion of semantic neighborhood of peers [2, 11, 13].

With respect to the previous approaches, the novel contribution of our work is related to the development of community aggregation techniques capable of combining ontological descriptions of peer interests with dynamic ontology matching techniques, to overcome the limitations of exact matching techniques adopted in most approaches and to provide semantic matchmaking capabilities in community formation, capable of dealing with dynamism and flexibility requirements of open networked systems. Furthermore, the use of ontology matching techniques provides a level of structuring of semantic neighbors of a peer for addressing a given query to the best matching peer(s) in the community.

7 Future work

In this paper, we have presented the work we are undergoing for semantic community formation in P2P systems. The work is at an initial stage of development and our future work will be in the following directions: semantic community management, semantic handshake techniques, and community-aware query propagation.

Semantic community management. We are interested in analyzing resilience and robustness properties of the semantic communities of peers. In particular, appropriate community management policies and techniques will be studied and devised for coping with the main events related to the life cycle of a committed community, such as the insertion of new members, the pruning of leaving participants, the unexpected peer failure, and the disband of the community. In response of insertion/pruning and failure events, the community is expected to react in order to re-arrange the communication structure of the community. For what concern the disband event, a timeout mechanism combined with a keep-alive technique could be adopted to realize the automatic disband of useless communities.

Semantic handshake techniques. For what concern the semantic handshake algorithm, we plan to implement such a semantic community aggregation protocol and to develop appropriate *commitment policies* for allowing a community founder to specify the requirements to be satisfied by the potential member peers for the establishment of an emerging community. Moreover, we are working on the definition of advanced consensus negotiation techniques in which the community ICard is the result of an active negotiation process in which the founder and the interested peers interact and discuss changes to the community ICard until an agreement among them is established. *Community-aware query propagation.* We intend to use simulation techniques for evaluating the performances of the community-aware query propagation at intra/inter community level and under different conditions of community overlap. Actually, in Helios, we have developed a basic semantic routing protocol which exploits ontology matching results to propagate a query to selected peers. However, such mechanism does not take into account communities [5]. We stress that, semantic communities emerge in consequence of user-driven events. This approach can foster the formation of a high number of small overlapping communities. For this reason, we are working on a clustering algorithm we have developed for information integration [4], and which is based on semantic reconciliation techniques in order to allow the aggregation of highly similar semantic communities and to allow the definition of an efficient structure for improving inter-community query propagation. Finally, we are interested in developing popularity-driven community aggregation techniques, where a peer founder can settle to advertise a semantic community on the basis of queries sniffed in the network. When a great number of queries in the network is due to similar requests, a peer can propose to found a semantic community regarding such a popular topic.

References

1. The HELIOS Project web site. http://islab.dico.unimi.it/helios/.
2. W.-T. Balke, W. Nejdl, W. Siberski, and U. Thaden. Progressive Distributed Top-k Retrieval in Peer-to-Peer Networks. In *Proc. of the 21st Int. Conference on Data Engineering (ICDE 2005)*, pages 174–185, Tokyo, Japan, April 2005.
3. M. Bonifacio, P. Bouquet, G. Mameli, and M. Nori. Peer - Mediated Distributed Knowledge Management. In *Proc. of Int. Symposium on Agent Mediated Knowledge Management (AMKM 2003)*, pages 31–47, Stanford, CA, USA, March 2003.
4. S. Castano, V. De Antonellis, and S. De Capitani Di Vimercati. Global Viewing of Heterogeneous Data Sources. *IEEE Transactions on Knowledge and Data Engineering*, 13(2):277–297, March/April 2001.
5. S. Castano, A. Ferrara, S. Montanelli, E. Pagani, G. P. Rossi, and S. Tebaldi. On Combining a Semantic Engine and Flexible Network Policies for P2P Knowledge Sharing Networks. In *Proc. of the DEXA GLOBE 2004 Workshop*, Zaragoza, Spain, September 2004.
6. S. Castano, A. Ferrara, S. Montanelli, and G. Racca. From Surface to Intensive Matching of Semantic Web Ontologies. In *Proc. of the DEXA WEBS 2004 Workshop*, Zaragoza, Spain, September 2004.
7. S. Castano, A. Ferrara, S. Montanelli, and G. Racca. Semantic Information Interoperability in Open Networked Systems. In *Proc. of the Int. Conference on Semantics of a Networked World (ICSNW 2004)*, Paris, France, June 2004.
8. J. Broekstra et al. A Metadata Model for Semantics-Based Peer-to-Peer Systems. In *Proc. of the WWW SemPGRID 2003 Workshop*, Budapest, Hungary, May 2003.
9. W. Nejdl et al. EDUTELLA: a P2P Networking Infrastructure Based on RDF. In *Proc. of the 11th Int. World Wide Web Conference (WWW 2002)*, Honolulu, Hawaii, USA, May 2002.
10. G.W. Flake, S. Lawrence, C.L. Giles, and F.M. Coetzee. Self-Organization and Identification of Web Communities. *IEEE Computer*, 35(3):66–70, March 2002.
11. P. Haase, R. Siebes, and F. van Harmelen. Peer Selection in Peer-to-Peer Networks with Semantic Topologies. In *Proc. of the Int. Conference on Semantics of a Networked World (ICSNW 2004)*, Paris, France, June 2004.
12. M. Khambatti, K. Dong Ryu, and P. Dasgupta. Structuring Peer-to-Peer Networks Using Interest-Based Communities. In *Proc. of the 1st Int. DBISP2P Workshop*, Berlin Germany, September 2003.
13. S. Staab, C. Tempich, and A. Wranik. REMINDIN': Semantic Query Routing in Peer-to-Peer Networks based on Social Metaphors. In *Proc. of the 13th Int. conference on World Wide Web (WWW 2004)*, New York, NY, USA, May 2004.