CAPSI' 2006 – 7ª Conferência da APSI

17 a 19 de Janeiro de 2007

Universidade de Aveiro, Aveiro, Portugal

Social Network Framework

Ricardo Manuel de Sousa Faria Marques Duarte

INESC-ID/Instituto Superior Técnico, Lisboa, Portugal

ricardo.duarte@inesc-id.pt

Alberto Manuel Rodrigues da Silva

INESC-ID/Instituto Superior Técnico, Lisboa, Portugal

alberto.silva@acm.org

CAPSI' 2006 – 7ª Conferência da APSI
17 a 19 de Janeiro de 2007
Universidade de Aveiro, Aveiro, Portugal

Social Network Framework

Abstract

Social networks are an emergent phenomenon on the Internet. It becomes increasingly necessary to be able to create models that can help us represent these structures, with appropriate focus on the underlying main concepts. A clear and formal definition of these concepts will allow both the explicitation of social networks, which already exist implicitly in several Information Systems today, and also the ability to integrate networks from different systems and from different domains. Moreover, an uniform representation of concepts allows the development and reutilization of algorithms specific to the social networking domain. In this text a Social Network Framework is presented to answer these needs. Focus is given on the conception of a modeling strategy to represent social networks. A discussion of model transformations and algorithms specific for social networks is also presented.

Palavras chave: social networks, modeling, integration, transformations, link mining.

1. Introduction

Social networks are emerging on the Internet as one of the most important recent phenomena in Information Society [Hempel 2005]. This kind of networks which have the main goal of capturing information about social relationships between people in digital support, enclose an intrinsic value, that can and should be explored in several different ways, ranging from direct marketing to the retrieval of people and extraction of information. These abilities are related to a more general potential that is inherent to social networks, which is the capability to express a context based on information about the neighborhood around a given individual in the network. Social networks also have the power to disclose the true identity of Internet users. In fact, when a link is established between two users, both are acknowledging and certifying the existence and identity of one another. This kind of identification will bring a greater degree of responsibility to the usage of information systems, because abusive behavior can be observed by the social network neighbors of each individual, and this will strongly relate digital responsibility to real life responsibility. Although the popularization of social networks is recent, these networks already exist more or less implicitly in all Information Systems which somehow support an online collaboration between people or organizational structures [SocialNetwork]. In this context, it becomes necessary to extract the information present in these systems, and to represent it in an explicit way, unraveling true social networks in a format which may be recognized by algorithms able to extract new information that was previously unknown from these networks. A new requirement that emerges from the growing proliferation of social networks throughout the Internet is the ability to integrate information captured from different social networks. Most social networks are specific to given domains or otherwise too general to contain all the relevant information needed to solve specific problems. The integration of networks from different domains may reveal new information which can be explored and analyzed. It is thus necessary to create integration mechanisms that allow us to infer new networks, with different relations and topologies, from existing social networks.

To answer the need for social networks analysis and integration, a Social Network Framework (SNF) is proposed in this paper. This framework encloses (1) the definition of a set of models to allow the representation of social networks, (2) importing mechanisms that will retrieve social networks from existing systems and represent them according to the defined models, (3) a set of transformations, allowing conversions between models, and (4) algorithms which may analyze the social networks to retrieve relevant information.

Figure 1 shows the package hierarchy of the Social Network Framework, and the existing dependencies between packages.

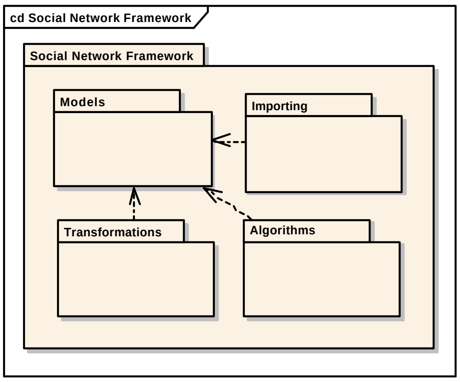

Figure 1 – Social Network Framework

In this Section we explained the motivation behind this work and proposed a Social Network Framework to deal with the identified problem. The remainder of this paper is divided into five additional sections: In Section 2, social network history is briefly introduced and an overview on previous work done on this area is given. In Section 3 a three layer modeling strategy to represent social networks is presented. We discuss different kinds of social network models intended for specific purposes: Specific Domain Models and a Canonical Model. In Section 4, some features that should be implemented to take advantage of the defined models are enunciated. Importing mechanisms to extract implicit social network from existing systems are exemplified, possible transformations between social network models are discussed, and some algorithms able to analyze social networks are listed. Finally, in Section 5 we present the main conclusions taken and discuss future work that can be done beyond the initial approach presented in this text.

2. State of the Art

Although the concept of "social networks" has been referenced for the first time in 1954 by J. A. Barnes [SocialNetwork], only in the mid-90s has this phenomenon found its firmed place in Information Society, through the large scale dissemination of the Internet. Before, the establishment of large social networks was difficult since it was hard to keep people constantly in contact with each others. It was also difficult to keep track of such networks. With the emancipation of the Internet, the natural human ability to build social relationships through communication has been bestowed, and some networks started to emerge based on websites, forums and mailing lists.

The multiplication of social networks may have started in 1995 with the website classmates.com [Classmates]. Since then, other networks have emerged, with a greater or lesser number of members: as an example, mySpace [MySpace] counts with more than 68.000.000 users, as of 2005. The list of websites supporting social networks continues to grow, counting with more general or more specific networks [SNList]. But social networks are not only present in internet websites: the implicit network in MSN Instant Messenger [IM] contact lists, in Google's Gmail [Gmail] invite-only e-mailing service, or in phone companies' call logs are possibly much larger than those contained in any particular website, and are already strongly related to real user identities.

In the scientific community, research in the area of social networks divides itself between systems for finding experts, such as ReferralWeb [Kautz 1997], whose objective is to explore the links between people in an organization, to quickly find experts on a given topic that can be reached through personal contact lists of people in the organization; agent-based computing paradigms [Wooldrige 2002]; or Link Mining [Getoor 2003], which cares about development of algorithms which can explore the relationships in linked structures, mostly focused on the hyperlinked structure of the World Wide Web.

3. Social Network Framework – Models

There are several ways in which the domain of social networks can be modeled. Although this domain is not inherently complex, some choices must be made regarding its representation. These choices will vary according to the purpose intended by the representation.

As a guide for taking such decisions, three general considerations have been taken into account: (1) our representation should be broad enough to allow the representation of most social networks implicit in existent applications; (2) the representation should be flexible to allow integration and transformation between different social networks; (3) the representation should allow a simple implementation, such that generic graph algorithms may be applied on it, with an acceptable performance, even within large scale sets of data.

3.1. Basic Concepts

A social network, as a representation of social relationships between people, encompasses two basic concepts: (1) Person, which represents the basic nodes of the network; (2) Relationship, which denotes the existence of some social relationship between person nodes.

The Person concept represents individuals, human beings, which will potentially display social behavior between them. Care must be taken when attempting to generalize the Person concept. Considering, for example, that an organization is a (collective) person, may be sometimes adequate and others not, but as a rule of thumb, such extensions should be avoided, or dealed with special care, because they may invalidate the rationale behind some decisions taken during the construction of the model here presented.

A social relationship, henceforth referred to only as relationship, represents some interaction between two or more people. Care must also be taken when considering which relationships may represent social interaction between people. A social relationship implies that all participants in the relationship are active in that relationship. For example, "I know George" will only be considered a social relationship, if George also knows me, and thus participates in the relationship. This does not invalidate that relationships may be asymmetrical, such as "Mary is John's mother".

3.2. Modeling Strategy

Social networks, in the real world, are complex, as there are many different kinds of relationships which may exist between people. Therefore, a social network model, which attempts to simultaneously represent multiple kinds of relationships between different sorts of people, must be complex to some extent. This complexity goes against the simplicity intended by our previously stated guidelines, yet, simply ignoring this complexity would result in the conception of a shallow model, with a very limited expressive power. The proposed solution to accommodate both simplicity and expressiveness is the creation of a set of models instead of one, each intended for a specific purpose, and then providing suitable mechanisms to transform ones into others.

Figure 2 shows the three layers architecture proposed for social network modeling. A Social Network Meta Model (SNMM) is defined to model the representation of Social Network Models (SNM), which model Social Network Instances (SNI). A SNI represents a set of relationships between actual people. The respective SNM represents all kinds of people and relationships allowed in that SNI. Finally the SNMM defines the possible relations between People Types and Relationship Types that can be represented in SNM's.

At the SNM layer we define the different models that will allow us to level the tradeoff between simplicity and expressiveness. To allow a simple representation we define a Canonic Model (CM). Instantiations of the CM are simple enough to be readily analyzed by generic graph algorithms, which is the CM's main purpose. Specific Domain Models (SDM) allow us to represent social networks from specific domains in our Social Network Framework. For each existing social network which we intend to integrate or analyze, a SDM should be defined. With appropriate transformations, the instances of different SDM's can be integrated into a more general SDM instantiation or converted into instances of the CM that can be analyzed by specific algorithms. These transformations are defined at the models level and then applied on actual instances.

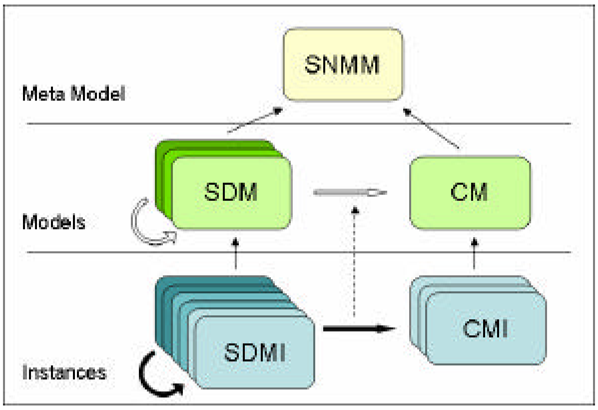

Figure 2 – Social Network Meta Model

Figure 2 shows the various models at their respective levels. Upward arrows crossing layers represent instantiations. Thick arrows represent transformation definitions (unfilled) and their respective instantiations (filled).

The SNMM and the CM are part of the core SNF models package. All SDM's should be developed as extensions to the framework, according to the particular social network domains we wish to use within the framework.

3.3. Social Network Meta Model

The Social Network Meta Model defines the possible elements and relations that can be expressed in Social Network Models. Figure 3 presents the SNMM. Instances of this meta model define social network types composed by a set of relationship types between given types of people, through specific role types.

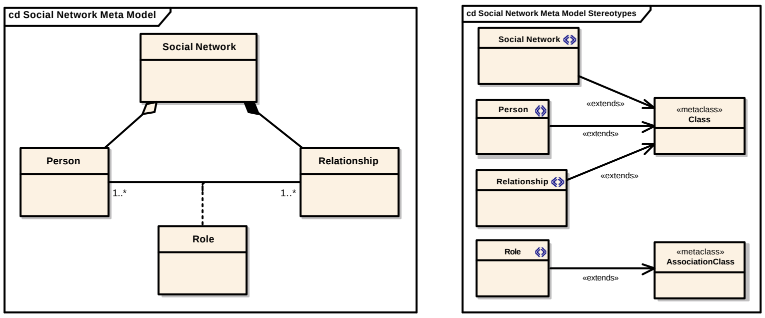

Figure 3 – Social Network Meta Model

3.4. Social Network Models

The SNMM allows us to represent Social Network Models, which define social networks as aggregations of relationships between people. These social networks also enclose the set of people who share relationships contained in the social network. A person not related to any other by a relationship, is not social by nature, and thus cannot be part of a social network. A person represents the identity of a real person, who can be part on several relationships which may pertain to different social networks. Roles allow people to participate differently in a relationship.

3.4.1 Specific Domain Models

Figure 4 shows, as example, an instantiation of the SNMM defining a SDM which represents a social network in a school domain. In this network there are two kinds of people, students and teachers, which relate between themselves through school class relationships.

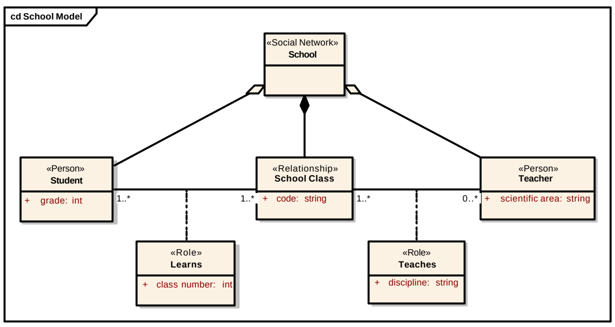

Figure 4 – School Specific Domain Model

An instantiation of the School SDM is shown on Figure 5, representing a simplified social network containing only one school class relationship between Teacher John Parker, and students, Paul, Susan and Mary.

3.4.2 Canonical Model

The Canonical Model represents a single view of a social network. This model allows only one specific kind of relationship to be represented, and is therefore a projection of the social network representing all existing relationships on a single dimension, which is the one regarding the relationship type of interest. The main purpose of the CM is to allow a simple representation that will be used to generically run graph algorithms, which may extract relevant information from the social network. As though, the representation for this model becomes very similar to the modeling of generic graph structures.

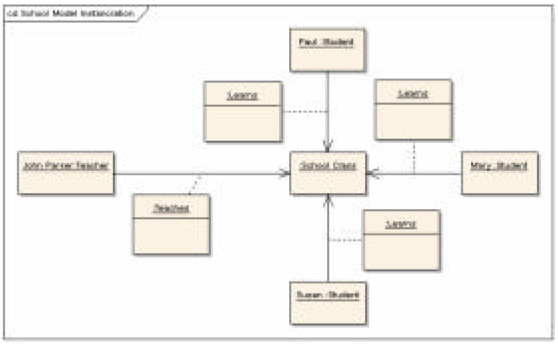

Figure 5 – School Model Instantiation

Figure 6 shows the proposed model for canonical social networks. As stated before, there is only one kind of relationships and each relationship associates only two people. A constraint should be enforced preventing that person who participates in the relationship through the source role is the same as the one assuming the target role.

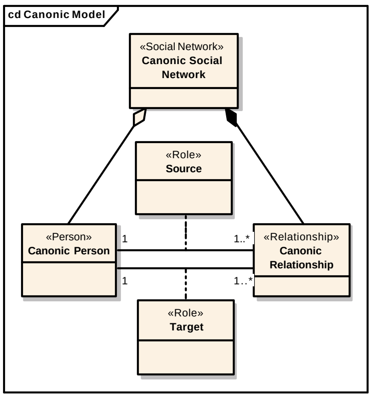

Figure 6 – Canonical Model for Social Networks

Each SDM may generate different CM's, one for each analytic purpose. Domain specific information is actually lost in the transformation process from SDM's to CM's, and so if we wish to analyze a network in some aspect of the problem domain, it is the transformation's responsibility to filter the generation of the canonic instantiation according to that semantic. For example, if we are interested in analyzing only classmate relationships in the 5th grade, the transformation should discard all teacher instances and filter students only to consider those from the desired grade. Some of the possible transformations are discussed in Section 4.2.

4. Social Network Framework – Features

To take advantage from our previously defined models, mechanisms should be implemented to (1) import social networks from existing systems, (2) specify transformations to integrate and convert models suited for different purposes, and (3) develop algorithms to extract previously unknown information from the represented social networks.

4.1. Importing

One of the main purposes of this work on social network modeling is that social networks that implicitly exist in actual systems may be expressed as true social networks. Developing a proper model to represent social networks, and allowing existing networks to be fit in this model, will allow us the take advantage of specific graph algorithms which can extract valuable information from the existing social networks. The SDM’s allows importing data from these systems into our social network framework. Once a given social network is represented as an instance of a SDM, it can be easily transformed into another type of model such as the CM, to take advantage of the SNF’s capabilities.

Figure 7 shows an example of a domain model of an existing learning management system [EscolaNaNet].

Instances of this domain model can be mapped into instances of the SDM defined in Figure 3. Importing teachers, students and school class relationships is straightforward from the entities with the same names in the domain model. Relations involving disciplines, school classes and teachers are mapped into teach roles, and registration entities are transformed into learn roles.

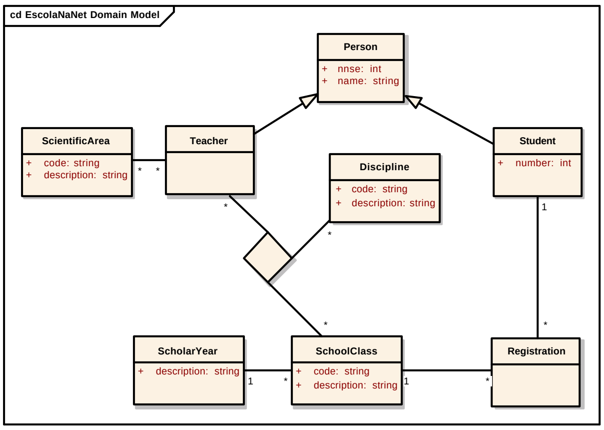

Figure 7 – Domain model of a learning management system in respect to the school class relationship.

4.2. Transformations

The representation we created of our school class SDM enables us to easily import data from an existing application, but in order to apply specific graph algorithms to extract information from the network, we need to transform our model into a canonic model, so that the algorithms can run effectively and efficiently. There are several possible transformations that we can apply to our models. We present the ones which we may use to convert general models into a canonic model, which tend to produce simpler models, but other transformations can also be built to produce more general models from simpler ones. Our proposed transformations are:

1. Projection - a social network model can be projected into a new one, by simply removing certain types of relationships, roles or persons. This way we can discard information which is not relevant to the solution of a specific problem, thus simplifying our model.
2. Selection - a transformation that allows us to obtain more specific networks from existing ones by filtering instances in the network according to a set of conditions enforced on attributes associated with the elements that may be defined in the model.
3. Equivalence Matching - It is possible to infer relationships based on the matching of topological units containing different relationships, roles and person types. In practice, this involves solving the problem of graph structure matching, whose solutions are computationally complex and generally don't scale well.
4. Canonicalization - the transformation of social networks into instances of the canonic model, entails two transformations: (1) simplification, which drops all the information expressed in the source model not contemplated in the canonic model, such as person types, roles and relationships' attributes, and (2) spanning, a transformation that combinatorially spans n-ary relationships into binary relationships.

The first three transformations are applied to convert SDM's into other more restrictive SDM's. The fourth converts a SDM into a CM. As our primary intent is to convert an existing SDM into a CM representation, we can first apply SDM to SDM transformations to produce intermediate models representing the subset of our domain that is pertinent for the intended analysis, and then apply the SDM to CM transformation that allows the algorithms to be applied.

4.3. Algorithms

Several generic graph algorithms exist which can be directly applied to retrieve information from the final canonic representation of our social networks. Some other algorithms specific to the domain of link analysis are especially appropriate to extract information from social networks.

Some generic graph algorithms can be applied being of practical interest in the context of social networks [Cormen 1997]:

1. Searching, which is one of the most common operations that we may want to apply on a social network. This may be breadth-first search or depths-first, although breadth-first search will be often more useful, since it looks first on individuals closer to the search root, which are often more important from a social network point of view.
2. Minimal Spanning Trees, which may be used to discover which individual should be chosen to more rapidly deliver a word-of-mouth message to an entire network. This can be interesting for marketing purposes, since a company could offer a product to clients with high network value, so that they will spread the word about it to others.
3. Shortest Paths, can reveal which is the more efficient way in which an individual may reach another. This algorithm may be useful to choose from the set of people retrieved by a search on a given criteria, the one which can be more easily contacted.
4. Strongly Connected Components, may be of some interest in finding communities. These communities are not strong, since we only know that people in the community are accessible to one another through their contact network, but may not be known to one another. Moreover, if relationship directionality is irrelevant in the social network, then the entire network will be a strongly connected component and in such cases the algorithm has no practical usefulness.

Algorithms from Link Analysis, an area pertaining to the more general research field of data-mining with the purpose of extracting new information on linked structures such as graphs, can be also applied on social networks. These include classification algorithms, which predict the value of nodes based on their links, and clustering algorithms, which attempt to group nodes of a same kind and are of special interest in finding communities on linked structures. Some examples follow:

1. Clustering in Directed Graphs – If our social network relationships are directional and if we believe that the nodes can be classified into two distinct categories such that nodes in one category typically have links to nodes in the other, we can apply algorithms to find these two distinct communities by partitioning the network in its two bipartite components. An example is the HITS algorithm [Kleinberg 1998], proposed by Kleinberg to find authority and hub pages on the hyperlinked structure of the Internet, which can be adapted to solve this problem, also retrieving the most important members in each discovered partition.
2. Clustering in Undirected Graphs – If our social network relationships' direction is irrelevant, maximum-flow algorithms may be used to identify communities conforming to the definition that a community member has more links to members inside the community than to outside members. Flake, Lawrence and Giles [Flake 2000] propose an algorithm to solve this problem, performing efficiently in large scale environments. This approach was proposed to find communities of websites, in order to identify communities of pages related to a topic defined by a starting set of chosen seed pages, but it has an even greater potential on the social network analysis field, since the proposed definition of a community fits better a real community of people than an artificial community of hyperlinked documents.
3. Clustering in Concentric Circles – Another approach on identifying communities in linked structure has been proposed by Zhou et al. [Zhou 2002]. This approach tries to identify communities at different scopes, defined by concentric circles around the some seed set of individuals, thus identifying communities with variable granularity.
4. Emergent Community Retrieval – Kumar et al. [Kumar 1999] discuss an algorithm to find communities in linked structures as soon as they start emerging on the network. This algorithm is based on crawling abilities to search for such emergent communities.
5. Viral Marketing Algorithms – One of the greatest potential of social networks is in identifying the network value of clients for marketing purposes, this activity is known as Viral Marketing [Domingos 2004]. Domingos et al. [Domingos 2001] propose an algorithm to assess this network value, based on markov random fields and bayesian theory.

5. Conclusions and Future Work

From the investigation discussed in this paper we conclude that much work on the area of social network is still open, and that this area has a growing potential which is now starting to be recognized and explored.

A Social Network Framework has been proposed to integrate and analyze existing social networks. Throughout this work, a set of models has been developed which are able to represent social networks extracted from real application scenarios, and mechanisms have been discussed to transform these models into a representation that is suitable to be understood by generic algorithms. We also conclude that several algorithms exist, both in graph literature and in the

Link Mining research field, which may be found useful and applicable in the context of social networking, to allow the extraction of the valuable information that these networks enclose.

The discussion presented in this paper opens some research paths that can be taken beyond the current investigation: (1) An interchange format should be discussed and defined in order to allow the representation and exchange of these models, so that integration from different kinds of systems can be made with a reduced effort. (2) The transformations discussed can be explored, and others may be found necessary. These transformations can be concretely defined and implemented, eventually evolving as a social network querying language or another mechanism that will implement the defined transformations. (3) Algorithms that perform over these networks should be implemented, refined and optimized to be of practical use in the context of social networks. (4) Moreover, real applications should be used to validate the presented models in real case scenarios. Integration can be obtained directly from application databases from specific systems, or by the use of crawling agents over the World Wide Web. (5) Privacy concerns must also be thoroughly evaluated, since they are essential in the assessment of what information can ethically and legally be extracted from social networks. (6) Finally, it would also be of interest the development of new systems, taking advantage of the defined model, transformations and algorithms, to support social network related tasks.

6. References

[Classmates] *Classmates.com*, classmate searching portal, http://www.classmates.com, Classmates Online, Inc. , (October 9, 2006) 1995-2006

[Cormen 1997] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, "Introduction to algorithms", MIT Press and McGraw-Hill Book Company, 18th edition, 1997.

[Domingos 2001] Pedro Domingos, et al., "Mining the network value of customers", In 7th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 57{66, ACM Press, 2001.

[Domingos 2004] Pedro Domingos, "Mining Social Networks for Viral Marketing", Department of Computer Science and Engineering, University of Washington, 2004

[EscolaNaNet] *EscolaNaNet*, http://www.escolananet.org, INESC-ID, (October 9, 2006) 2006

[Flake 2000] Gary William Flake, Steve Lawrence, C. Lee Giles, "Efficient Identification of Web Communities", in Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD-2000), 2000

[Getoor 2003] Lise Getoor, "Link Mining: A New Data Mining Challenge", ACM SIGKDD Explorations Newsletter archive, Volume 5, Issue 1 column: Position papers on MRDM, 2003

[Gmail] *Gmail*, e-mailing service, http://www.gmail.com, Google, (October 9, 2006) 2006.

[Hempel 2005] Hempel, J., Lehman, P. "The MySpaceGeneration", cover story in Business Week Magazine, December 12, 2005

[IM] *MSN Instant Messenger*, messaging service, http://messenger.msn.com, Microsoft, (October 9, 2006) 2006

[Kautz 1997] Henry Kautz, Bart Selman, Mehul Shah, "Referral Web: Combining Social Networks and Collaborative Filtering", AT&T Laboratories, in Communications of ACM, vol.40, nº3, March 1997.

[Kleinberg 1998] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", in Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998

[Kumar 1999] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, "Trawling the web for emerging cybercommunities", in Proceedings of The 8th International World Wide Web Conference, 1999

[MySpace] *MySpace*, personal webspaces portal, http://www.myspace.com, MySpace.com, (October 9, 2006) 2003-2006

[SocialNetwork] *Social Network*, http://en.wikipedia.org/wiki/Social_network, Wikipedia, (May 3, 2006) 2006

[SNList] *List of Social Network Websites*, list of the largest social network websites, http://en.wikipedia.org/wiki/List_of_social_networking_websites, Wikipedia, (May 3, 2006) 2005

[Wooldrige 2002] Michael Wooldridge, "An Introduction to MultiAgent Systems", John Wiley and Sons, ISBN 047149691, 2002

[Zhou 2002] Wei-Jun Zhou, Ji-Rong Wen, Wei-Ying Ma, Hong-Jiang Zhang, "A Concentric Model for Community Mining in Graph Structures", 2002