Social networks and statistical relational learning: a survey

Floriana Esposito*, Stefano Ferilli,
Teresa M.A. Basile and Nicola Di Mauro

Department of Computer Science,
LACAM Laboratory,
University of Bari 'Aldo Moro',
Via E. Orabona, 4 – 70125 Bari, Italy
E-mail: esposito@di.uniba.it
E-mail: ferilli@di.uniba.it
E-mail: basile@di.uniba.it
E-mail: ndm@di.uniba.it
*Corresponding author

Abstract: One of the most appreciated functionality of computers nowadays is their being a means for communication and information sharing among people. With the spread of the internet, several complex interactions have taken place among people, giving rise to huge information networks based on these interactions. Social networks potentially represent an invaluable source of information that can be exploited for scientific and commercial purposes. On the other hand, due to their distinguishing peculiarities (huge size and inherent relational setting) with respect to all previous information extraction tasks faced in computer science, they require new techniques to gather this information. Social network mining (SNM) is the corresponding research area, aimed at extracting information about the network objects and behaviour that cannot be obtained based on the explicit/implicit description of the objects alone, ignoring their explicit/implicit relationships. Statistical relational learning (SRL) is a very promising approach to SNM, since it combines expressive representation formalisms, able to model complex relational networks, with statistical methods able to handle uncertainty about objects and relations. This paper is a survey of some SRL formalisms and techniques adopted to solve some SNM tasks.

Keywords: statistical relational learning; SRL; social network modelling; social network analysis.

Reference to this paper should be made as follows: Esposito, F., Ferilli, S., Basile, T.M.A. and Di Mauro, N. (2012) 'Social networks and statistical relational learning: a survey', *Int. J. Social Network Mining*, Vol. 1, No. 2, pp.185–208.

Biographical notes: Floriana Esposito is a Full Professor of Computer Science at the University of Bari, and since 1989 chairs the Laboratory for Knowledge Acquisition and Machine Learning (LACAM). Her research interests, initially in the field of statistical pattern recognition, moved to the area of artificial intelligence and machine learning, concerning the logico-algebraic foundations of inductive learning, the integration of numerical and symbolic methods, the computational models of conceptual learning. She has been in the directorial board of the Italian Association for Artificial Intelligence and responsible of the Machine Learning Group. In 2006, she was elected as a European Coordinating Committee for Artificial Intelligence Fellow (ECCAI).

Stefano Ferilli is Associate Professor of Computer Science and Head of the Inter-departmental Center for Logics and its Applications (CILA) at the University of Bari, Italy. His research interests include logic and algebraic foundations of machine learning, theory revision, inductive logic programming, multi-strategy learning, and knowledge representation.

Teresa M.A. Basile received her Laurea degree in Computer Science in 2001 and PhD in Computer Science in 2005 from the University of Bari, Italy. Since April 2005, she is a researcher at the Department of Computer Science, University of Bari. Her research interests include symbolic machine-learning techniques, with special focus on the cooperation of different inferences strategies in incremental learning frameworks, and their application to document classification and biological/medical data understanding.

Nicola Di Mauro is an Assistant Professor of Computer Science at University of Bari, Italy. He received his PhD in Computer Science from University of Bari and Laurea degree from the same university. His research activity mainly concerns machine learning, statistical relational learning and probabilistic inductive logic programming.

1 Introduction

After their beginning as computation-focused machines, and a next stage involved in data-processing, computers have been characterised in the last decades as the main means for communication and information sharing among people all over the world. Supported by the World Wide Web (WWW), this perspective has taken the form of several (often huge) groups of (various kinds of) information items linked to each other in different ways, which are referred to as *information networks*. The WWW itself is an example, where the items are hypertextual documents connected to each other by hyperlinks. *Social networks* has flourished in the last years providing growing amounts of data, leading to the urgent need for *social network mining* (SNM) methods and techniques able to analyse them in order to gain useful, high-level information that emerges from the overall network and cannot be drawn considering single items separately.

With the spread of the internet, lots of people have started contributing to build and extend these networks, by subscribing, accessing and exploiting several kinds of websites on the WWW (and on its evolution known as the Web 2.0) and services. This gave rise to huge networks, such as e-mail communication networks, instant messenger networks, mobile call networks, and friends networks. A few outstanding examples are:

- citation networks, concerned with storing scientific papers, and with relating these papers and their subjects through their authors and the co-authorship or mutual reference relationships (e.g., DBLP)
- semantic network service (SNS) websites, basically organised according to people and their mutual friendship or professional connection (e.g., Facebook, Linkedin, UNYK)
- social shopping websites, focused on e-commerce (e.g., Amazon) and opinion sharing about products (e.g., Dooyoo)

- social media websites, that provide suggestions about music, movies, etc. based on user tastes and typical behaviours (e.g., Last.fm).

Interest in social network analysis is motivated by several possible tasks, aimed at extracting information at different levels of granularity, ranging from the whole network up to single items. For instance, one might want to extract information about a person or an object that is not explicit in its description, but emerges from general considerations derived from its direct or indirect relationships in the network; or one might be interested in emerging groups of elements having similar behaviour or similar tastes, as determined by their features in the network. Thus, possible outcomes of the analysis activity are the discovery of social structures, social position or role of individuals.

The peculiar feature distinguishing social networks is their consisting of rich collections of objects linked into complex relational networks. This makes SNM quite different from other information extraction tasks, and requires the exploitation of new techniques for carrying it out. First-order logic is the typical setting that provides sufficiently powerful representation languages to handle relationships. A machine learning subfield able to deal with logical representations is inductive logic programming (Muggleton, 1991), where both instances and learned models are represented using logic programming. Inductive logic programming aims to find a hypothesis H (a logic program) from a set of positive and negative examples fulfilling the constraint that the hypothesis H logically explains all positive examples and none of the negative examples. In order to have sufficiently robust and efficient systems able to deal with large quantities of possibly noisy data a lot of work, known under the names of *statistical relational learning* (SRL) (Getoor and Taskar, 2007), or probabilistic inductive logic programming (PILP) (De Raedt et al., 2008b), is appeared in the last few years. Models belonging to SRL and PILP combine expressive representation formalisms with statistical methods to perform probabilistic inference and learning on relational networks.

In this paper, we survey some SRL formalisms and techniques for social network modelling and analysis. After introducing the setting and tasks of social network analysis in Section 2, and providing a brief overview of the fundamental of SRL in Section 3, we present the application of several SRL techniques to different tasks in social network analysis in Section 4, before concluding the paper in Section 5.

2 Social networks analysis

Social network analysis regards the study of relations among individuals, represented as a network, aimed at analysing aspects such as social structures and role analysis. A (social) network is normally represented with a complex structured graph where each potential node/edge is considered a random variable describing the state of the node/edge. From a machine learning point of view, the task is to learn the probabilistic dependencies between these random variables. In a graph representation, nodes represent the actors involved in the network, while edges denote the connections among the actors. Many relevant tasks can be setup on social networks; those relevant to the machine learning field are (Getoor, 2003; Getoor and Diehl, 2005; Tang and Liu, 2010):

- *link prediction*, concerned with identifying when two actors may be connected or, more importantly, whether they may be connected in the future

- *community detection*, that tries to detect communities (groups of actors) by studying the network structure and topology
- *object classification* and outlier detection, whose goal is to correctly predict the labels that may be associated to the actors
- *position/role analysis*, aimed at identifying the role associated with different actors
- *information diffusion* and *viral marketing*, that study and model how the information propagates in the social network, resulting in emerging relevant trends that can be exploited for a deeper understanding of the network and its elements.

The link prediction task is very important in information network analysis. Link inference and link prediction are two recent statistical machine learning problems appeared as a result of the increasing interest in the broader problem of *link mining* in social networks [see Getoor and Diehl (2005) and Senator (2005) for a survey]. In a *static perspective*, it consists of identifying the connections between two items in a network, even if a direct connection between them is not explicitly present in the network graph. Inferring such an implicit link would allow to apply similar actions on those items, e.g., submitting them similar items of interest in a library or e-commerce context. In an *evolutionary* perspective, link prediction consists of the task of finding, given a snapshot of the network, which unobserved links among items are likely to occur in the future. This is a crucial topic in information network analysis, since effective predictions might allow to foresee how the whole net evolves, and hence to better understand and handle the context they represent. In both cases, it needs to work on descriptions that are not limited to simple attribute-value pairs, but involve relationships as a non-negligible component. As an additional issue, real world networks, such as social networks, are characterised by extremely noisy and sparse data.

After initial efforts focused on unsupervised learning, several works started to explore supervised approaches to learning models for link prediction (Lichtenwalter and Chawla, 2011) implemented a scalable and efficient multi-core tool link prediction including both unsupervised and supervised techniques), and on assessing which features can be more predictive. Hasan et al. (2006) identify key and efficient features to be exploited by propositional statistical learning algorithms. Other works exploit a random walk approach, guided towards promising nodes to be linked by node (Liu and Lu, 2006) or edge-related (Backstrom and Leskovec, 2011) functions. Also Liben-Nowell and Kleinberg (2007) define *a proximity* measure, taking different similarity measures (based on shortest path, node neighbourhoods, ensemble of all paths, coupled with meta-approaches that exploit their results for higher-level selection of best links) as indicators of the likelihood of introducing the corresponding edge. Lichtenwalter et al. (2010) focus on sparse networks, where only a few potential links actually form, and present an effective flow-based predicting algorithm and a completely general framework that outperforms unsupervised link prediction methods. Other works are based on the non-parametric Bayesian framework: Cao et al. (2010) address the data sparsity problem using a *collective link prediction* approach, which jointly predicts different kinds of links; Miller et al. (2009) use non-parametric *latent feature* models to simultaneously infer both the number of features and which entities have each feature.

Several graph-based approaches have been attempted to solve the link prediction problem. Recently, Sarkar et al. (2011) correlate graph generation models associated to a

latent metric space and link prediction heuristics to study several indicators such as role of node degree, path length, and non-determinism in the link generation process. Acar et al. (2009) exploit matrix and tensor-based methods to predict links in bipartite graphs that change in time. Leroy et al. (2010) face the *cold start* link prediction problem (in which the structure of the network is missing and only information regarding the nodes is available) by first generating an implicit network in the form of a probabilistic graph and then applying probabilistic graph-based measures to produce the prediction.

The identification of communities in large networks is a very important problem because it allows to compactly represent and manage elements by class, and to notice trends in these groups. If the network is considered as a graph, detecting a community means finding a sub-graph whose elements exhibit strong linkedness and high similarity according to some parameter. Since many problems on graphs are computationally intractable, or practically infeasible for large inputs (as in the case of social networks), it turns out that community identification is very complex as well. Community detection approaches can be distinguished, from the perspective of their resulting output, based on their allowing or not overlapping communities (in the former case, a node can belong to several communities; in the latter, each node belongs to at most one community). In fact, this task is strictly related to clustering: fixed an objective function that allows to find sets of nodes with dense connections within sets and sparse connections between sets, approximation algorithms or heuristics can be applied to find the clusters accordingly. Leskovec et al. (2010) evaluate and compare a range of methods and objective functions, and examine several classes of approximation algorithms to optimise such functions, also considering community size as a quality indicator. To obtain a scalable algorithm that considers the whole network, Gargi et al. (2011) propose to combine a pre-processing stage, a local clustering stage, and a post-processing stage to generate labelled and consistent clusters of YouTube videos. Hohwald et al. (2009) work on mobile phone calls; they allow overlapping communities and focus on networks with unobserved interactions (missing edges) to predict future interactions. The methods also models inter and intra community interactions and their strength, used to decide whether merging communities based on binary decisions rather than weights. After applying an agglomerative hierarchical technique, links between communities are searched by using artificial neural networks and logistic regression, which perform better than a naive majority-class classifier.

Maiya and Berger-Wolf (2010) do not allow overlapping communities, and recast the problem to univariate collective inference by defining the *expansion sampling* method, that allows to work on sampled items only. Random samples are considered representative of the overall network according to maximum expansion factor, where the maximum expander set can be approximated based on a greedy algorithm using snowball sampling or on Markov chain Monte Carlo simulation.

Lozano et al. (2006) analyse large social datasets using a methodology based on community division, that allow to mix link and node attribute information. Hui et al. (2007) propose distributed approaches for mobile devices to detect both static and temporal communities that can approximate their corresponding centralised methods. Leung et al. (2009) analyse and extend an existing label propagation algorithm to obtain real-time community detection. Meo et al. (2011) present an approach that exploiting a novel measure of edge centrality discovers the community structure adopting a strategy inspired by the Louvain method, efficiently maximising the network modularity.

The identification of noteworthy elements in a social network may allow a deeper understanding of the behaviour of such items that could not be obtained without reference to the network as a whole. In particular, outliers are items showing a strange behaviour with respect to the overall network. Gao et al. (2010) detect *global* outliers (determined by their intrinsic features only) and *contextual* ones (that are 'abnormal' only relatively to related elements) during community discovery, based on an integrated probabilistic model, where generative mixture models are used to describe items, and hidden Markov random fields are used to determine the joint distribution of both data and links. Conversely, Tang and Liu (2011) ignore node features and leverage their relationships only, assuming that different kinds (labels) of links describe latent social dimensions. They use support vector machines and logistic regression to build from labelled nodes a discriminant classifier based on these dimensions (where each item may belong to many classes), and then apply it to the unlabeled ones.

Karamon et al. (2007) define primitive operators for structural feature generation, whose combination automatically yields several social network indexes (some well-known, some others new), this way bridging the gap between the aggregation of network features for relational data mining and traditional analytical methods for social network analysis. Akoglu et al. (2009) propose a scalable algorithm, defining features and rules that are useful to identify nodes with strange behaviour in weighted graphs. Aggarwal et al. (2011) use a structural connectivity model for defining outliers in massive network/graph streams, dynamically partitioning the network to handle the sparsity problem and designing a reservoir sampling method to maintain structural summaries of the network.

Chakrabarti (2004) works on the cross-point between clustering and outlier detection, relying on information theoretic principles to provide a parameter free and scalable algorithm aimed at overcoming the problems of other methods such as *k*-means clustering, METIS graph partitioning and singular value decomposition or principal component analysis. Bilgic et al. (2007) jointly face object classification and link prediction in case of missing or wrong attributes and links, proposing a collective algorithm that interleaves the two tasks.

Social roles played by people in their interaction get a peculiar importance in online systems, because different kinds of roles can be associated to different types of users and, as a consequence, allow to generalise users' behaviour and to detect and manage communities. Social roles in online community are defined in Gleave et al. (2009) as a combination of social psychological, social structural, and behavioural attributes; they also provide measurement and analysis strategies for identifying them in Usenet and Wikipedia.

Ouimet et al. (2004) propose an approach to construct sociometric variables measuring the network positions of firms in a small industrial cluster: among the network measures used (degree, betweenness and effective size), only degree and effective size are positively correlated with radical innovation. McCallum et al. (2005) present the author-recipient-topic (ART) model, which learns topic distributions according to the relationships between people and is able to predict people's roles on e-mail corpora. Hsu et al. (2008) investigate in the framework of first-order logic whether a concept can be defined using social positions, showing that this sometimes implies its definability using the corresponding social positions.

Possible applications of diffusion analysis range from commerce, to medicine, to sociology. Viral marketing, in particular, focuses on how recommendations or just

opinions on given products, provided by some elements of the network, affect the adoption of those products by other (directly or indirectly related) elements of the network.

Buskens and Yamaguchi (1999) focus on the efficiency of information transfer, proposing a model in which nodes may retain information after sending it that is able to predict diffusion times at different granularity levels and to generalise several network measures. They also analyse the relationships between diffusion times and centrality measures according to a series of network measures. Lafferty and Lebanon (2002) introduce a family of kernels for statistical learning, providing a natural way of combining generative statistical modelling with non-parametric discriminative learning, proving theoretical results and experiments on text classification. Kempe et al. (2003) provide a general approach and specific results about approximation guarantees on efficient and effective greedy algorithms for the NP-hard optimisation problem of selecting the most influential nodes, based on sub-modular functions.

Gruhl et al. (2004) study the dynamics of information propagation in personal publishing environments, presenting both a macroscopic (at the network level) and a microscopic (at the level of individuals) characterisation of propagation, based on the theory of infectious diseases, deriving an algorithm to induce the propagation network from a sequence of posts. Yang and Leskovec (2010) note that patterns of influence depend on the type of the node and the topic of the information, developing a linear influence model that determines the number of newly infected nodes as a function of the nodes previously infected, and becomes scalable for large datasets in its non-parametric formulation (reducing to a simple least squares problem). Romero et al. (2011) study how tokens (hashtags) having different topic spread on a Twitter network depending on their 'stickiness' and 'persistence', also based on their initial adopters and of the related subgraph structure.

3 Statistical relational learning

The vast interest in SRL (Getoor and Taskar, 2007) and in PILP (De Raedt et al., 2008b) has resulted in a wide variety of different formalisms, models and probabilistic programming languages (De Raedt et al., 2008a).

Probability logic-based formalisms define probabilities using either a *direct* or an *indirect* approach (Cussens, 2007). In the former, probabilities are explicitly provided for each probabilistic fact, and the corresponding model is closely related to a Bayesian network. Formalisms falling into this category are probabilistic horn abduction (PHA) (Poole, 1993), probabilistic logic programming (PLP) (Ng and Subrahmanian, 1992), relational Bayesian networks (RBNs) (Jaeger, 1997), Bayesian logic programming (BLP) (Kersting et al., 2000), stochastic logic programmes (SLPs) (Muggleton, 1996), PRISM (Sato and Kameya, 1997), CLP(BN) (Costa and Cussens, 2003), ProbLog (De Raedt et al., 2007), and logic programmes with annotated disjunctions (LPADs) (Vennekens et al., 2004). Since some of these languages can be translated into Bayesian networks, when the networks contain hidden variables, learning the parameters of these languages requires the use of techniques for learning from incomplete data such as the expectation maximisation (EM) algorithm (Dempster et al., 1977) or the recent relational information bottleneck (RIB) framework (Riguzzi and Di Mauro, 2012). In the indirect approach,

conversely, formulæ are not explicitly associated to their probability, and the probability of a possible world is defined in terms of its features by means of an associated real-valued parameter. A formalism falling in this category is Markov logic networks (MLNs) (Richardson and Domingos, 2006).

These approaches define a probability distribution in a logic-based formalism and solve the so-called *inference* problem, consisting in the computation of probabilities to answer specific queries. Their logical interpretation is in terms of classical least Herbrand models, while the probabilistic semantics is in terms of a *possible worlds* semantics. Since learning is a fundamental component of the systems based on SRL formalisms, there are two additional problems that should be considered: *parameter estimation* and *structure learning*. The assumption is that the observed data are a sample generated from an unknown distribution, and that the aim is learning such a distribution. When the structure of the model is known, there is a need to learn the parameters of the model. In general, both the structure and its parameters must be learned.

Another category of modelling approaches concerns the combination of relational database models and graphical models as for the following formalisms: probabilistic relational models (PRMs) (Friedman et al., 1999), probabilistic entity relational models (PERMs) (David Heckerman and Meek, 2007) and relational Markov networks (RMNs) (Taskar et al., 2007).

The parametric approaches to SRL listed above focus on probabilistic models with finitely many parameters, selecting a single model that performs best. Other non-parametric approaches work with probabilistic models with infinitely many parameters such as infinite (hidden) relational models (IHRMs) (Kemp et al., 2006; Xu et al., 2006) and the multi-relational Gaussian process model able to deal with an arbitrary number of relations recently proposed in Xu et al. (2009).

Even if SRL formalisms are able to deal with complex domains their parameter learning and structure learning algorithms are not efficient when compared to classical propositional statistical learning methods. Recent works on *lifted inference* (Poole, 2003; de Salvo Braz et al., 2005) allows expressive representations whose inference is made much cheaper by abstracting away from specific instances of random variables and dealing instead with whole classes thereof at once.

Another possible perspective towards SRL consists in restricting expressiveness, this way allowing for more efficient learning and inference algorithms. This category includes, among others, the following recent non-parametric approaches: nFOIL (Landwehr et al., 2005), that integrates the naive Bayes probabilistic model with a relational rule learner, kFOIL (Landwehr et al., 2006), where a relational kernel function is learned and defined in terms of a small set of interpretable relational features, Lynx (Di Mauro et al., 2011), that combines the naive Bayes probabilistic model with relational query construction and selection, and rsLDA (Taranto et al., 2011b) that combines a relational feature construction approach with the latent Dirichlet allocation hierarchical Bayesian model.

In the following, we briefly introduce two examples of SRL formalisms, Bayesian logic programmes (Kersting and De Raedt, 2007) extending the direct graphical model of Bayesian networks, and MLNs (Richardson and Domingos, 2006) extending the undirected graphical model of Markov networks. In order to describe these SRL formalisms we firstly report some basic definitions about first-order logic.

A first-order logic alphabet consists of a set of *constants*, a set of *variables*, a set of *function* symbols, and a non-empty set of *predicate* symbols. Each function symbol and

each predicate symbol have an associated number (its arity) specifying how many terms it must be applied to. A *term* is defined to be a constant symbol, a variable symbol, or an n-ary function symbol applied to n terms. An *atom* is a predicate symbol of arity n applied to n terms. Clauses are formulas of the form $A \leftarrow B_1, \ldots, B_m$, representing implications $(B_1 \wedge \ldots \wedge B_m) \Rightarrow A$), where A and the B_i's are atoms (in particular, A is called the *head* of the clause, and the B_i's its *body*) and all variables are implicitly understood to be universally quantified. A *logic programme* consists of a set of clauses. A term, atom or clause is called *ground* when there is no variable occurring in it. A substitution is an assignment of terms to variables. Applying a substitution θ to a set of atoms c, denoted as $c\theta$, provides the instantiated set of atoms where all occurrences of the variables are simultaneously replaced by the corresponding term. The *Herbrand base* of a logic programme is the set of all ground atoms that can be built on the predicate, constant and function symbols in the alphabet of the programme. A *Herbrand interpretation* for a logic programme is a subset of its Herbrand base, while its *least Herbrand model* consists of all facts belonging to the Herbrand base and logically entailed by the programme.

3.1 Bayesian logic programming

Here, we describe *Bayesian logic programmes* (Kersting and De Raedt, 2007) by firstly introducing the graphical model that provides its foundation, i.e., Bayesian networks (Pearl, 1991).

A Bayesian network is a directed acyclic graph G representing a dependency structure over a set of random variables $X = \{X_1, \ldots, X_n\}$, where each variable X_i is represented by a node in the graph and an edge between two variables denotes their direct influence. Furthermore, the random variables in X are associated to corresponding domains $D = \{D_1, \ldots, D_n\}$ and probability distributions $P = \{P_1, \ldots, P_n\}$ with $P_i = P(X_i \mid \mathrm{Pa}_{X_i})$, where Pa_{X_i} denotes the set of parents (i.e., the direct predecessors) of X_i in G. A Bayesian network represents a probability distribution $P(X_1, \ldots, X_n)$ over the variables in X, and because of the conditional independence assumption of Bayesian networks (i.e., $P(X_i \mid A, \mathrm{Pa}_{X_i}) = P(X_i \mid \mathrm{Pa}_{X_i})$), we can write:

$$P(X_1, ..., X_n) = \prod_i P(X_i \mid \mathrm{Pa}_{X_i}).$$

Bayesian logic programmes (Kersting and De Raedt, 2007) unify Bayesian networks with logic programming allowing to overcome both the propositional nature and limitation of Bayesian networks and the purely logical nature of logic programmes.

The concept of logical clause is extended to that of Bayesian clause that is an expression of the form $A \mid A_1, \ldots, A_n$ where $n \geq 0$, the $A, A_1, \ldots, A_n$ are Bayesian atoms, which means that they have an associated finite set of possible states, and all Bayesian atoms are universally quantified. When $n = 0$, the clause is called a Bayesian fact and is expressed as A. Assuming that for each Bayesian predicate there is a corresponding combining rule (i.e., a function mapping finite sets of conditional probability distributions $\{P(A \mid A_{i1}, ..., A_{in_i}) \mid i = 1, ..., m\}$ onto one combined conditional probability distribution $P(A \mid B_1, \ldots, B_k)$ with $\{B_1, ..., B_k\} \subseteq \bigcup_{i=1}^{m} \{A_{i1}, ..., A_{in_i}\}$, it is possible to formally define a Bayesian logic programme as consisting of a (finite) set of Bayesian clauses. For each

Bayesian clause there is exactly one conditional probability distribution and for each Bayesian predicate there is exactly one combining rule.

A declarative semantics for Bayesian logic programmes can be formalised using the *annotated dependency graph* (Kersting and De Raedt, 2007). The dependency graph is a directed graph whose nodes correspond to the ground atoms in the least Herbrand model $LH(B)$. It encodes the direct influence relation over the random variables in the least Herbrand model. In particular, there is an edge from a node X to a node Y if and only if there exist a clause c and a substitution θ such that $Y = head(c\theta)$, $X \in body(c\theta)$ and for all ground atoms Z in $c\theta$: $Z \in LH(B)$.

A Bayesian logic programme is a template for a Bayesian network whose nodes are the relevant random variables. Traditional techniques used for parameter estimation of Bayesian networks, such as the EM algorithm, can be adapted for learning the parameters of Bayesian logic programmes. As regards the structure learning of Bayesian logic programmes, classical refinement operators used in inductive logic programming can be used.

3.2 Markov logic networks

This section reports an extension to the logical case of Markov networks model. A *Markov network* is an undirected graphical model for the joint distribution of a set of variables, made up of an undirected graph, containing a node for each variable, and a set of potential functions for each clique in the graph. Given a set of variables X and a set of potential functions ϕ_k, a Markov network represents the following joint distribution:

$$P(X = x) = \frac{1}{Z}\prod_k \phi_k\left(x_{\{k\}}\right),$$

where $x_{\{k\}}$ represents the state of the k^{th} clique, and Z is the *partition function* given by $Z = \sum_{x \in X}\prod_k \phi_k\left(x_{\{k\}}\right)$. A Markov network can be also represented in a log-linear way as

$$P(X = x) = \frac{1}{Z}\exp\left(\sum_j w_j f_j(x)\right),$$

where each potential function has been replaced by a weighted sum of features.

MLNs (Richardson and Domingos, 2006) extend Markov networks to first-order logic, where a *possible world* is an assignment of truth values to all possible groundings of predicates. A first-order formula can be seen as a hard constraint on the set of possible worlds (if a world violates even one formula, it has zero probability), while MLNs soften this constraint (when a world violates one formula it is less probable, but not impossible). To this purpose, each formula has associated a weight representing how strong a constraint it is. Thus, a MLN is a set of pairs (F_i, w_i), where F_i is a formula in first-order logic and w_i is a real number.

Given a set of constants C, a MLN defines a Markov network containing one binary node for each possible grounding of each predicate appearing in the MLN, whose value is 1 if the ground atom is true, or 0 otherwise. Furthermore, the Markov network contains one feature for each possible grounding of each formula F_i in L. The value of this feature is 1 if the ground formula is true, or 0 otherwise. The weight of the feature is the w_i

associated with F_i in L. The probability distribution over possible worlds x specified by the ground Markov network is given by

$$P(X = x) = \frac{1}{Z}\exp\left(\sum_i w_i n_i(x)\right) = \frac{1}{Z}\prod_i \phi_i\left(x_{\{i\}}\right)^{n_i(x)},$$

where $n_i(x)$ is the number of true groundings of F_i in x, $x_{\{i\}}$ is the state of the atoms appearing in F_i, and $\phi(x_{\{i\}}) = e^{w_i}$. An MLN can be viewed as a template for constructing Markov networks.

The parameters of an MLN may be learned using many techniques to optimise the pseudo-log-likelihood, while its structure is usually learned adopting methods borrowed from inductive logic programming.

4 Social network analysis with SRL

Social networks are usually represented by a complex relational network involving many linked objects. SRL combines expressive knowledge representation formalisms with statistical approaches resulting in a good choice to perform probabilistic inference and learning on relational networks. In this section, we provide a survey of works that adopt SRL formalisms to solve specific social network tasks.

4.1 Object classification

Object classification regards the problem of predicting the label of an object within the network using its observed attributes and the observed/unobserved labels/attributes of the objects in its neighbourhood. To solve this task, it is necessary to perform *collective classification* (Jensen et al., 2004) where relationships among objects must be taken into account in order to enhance the predictive accuracy of the model: the labelling of an object should depend on the labels of its neighbours. Collective classification is one of the main tasks closely related to SRL as already reported in Chakrabarti et al. (1998), Taskar et al. (2001), Neville et al. (2003) and Jensen et al. (2004). In a social network, where nodes represent actors, the actor-actor links are used to boost the accuracy of local classifiers or even provide classification labels in the absence of local features (Zheleva et al., 2010). SRL techniques adopted to solve this problem assume that knowing the label of a particular object of the network can correctly guide the inference of the other nodes' labels. Methods adopting a collective classification approach can significantly outperform classification methods that ignore the relationships between nodes in the network (Sen et al., 2008). An interesting study has been reported in Macskassy and Provost (2007) that introduces a network learning toolkit that enables in-depth studies of techniques for SRL and classification with networked data.

Collective classification corresponds to a combined classification of interlinked objects using correlations between the label of an object and its observed attributes, the correlations between the label of an object and the observed attributes/labels of objects belonging to its neighbourhood, and the correlations between the label of an object and the unobserved labels of objects belonging to its neighbourhood. One commonly used method for collective classification is the *iterative classification algorithm* (ICA).

Assuming to have a local classifier, ICA predicts the best label for a node y_i taking the values of all other nodes in its neighbourhood. Since, some values in the neighbourhood might be unknown, the label of y_i is predicted by estimating all the values in the neighbourhood, and the process is iteratively repeated until the assignments to the labels get stable.

Taskar et al. (2002) proposed the use of a joint probabilistic classification model for a collection of related entities introducing the framework of RMNs that compactly defines a Markov network over a relational dataset. They extended the previous work (Taskar et al., 2001) on PRMs for collective classification overcoming its limitation in some domains where cycles in the link graph lead to cycles in the corresponding Bayesian network. Experimental results on web page classification proved the validity of the proposed model for modelling relational dependencies among entities. Following Taskar et al. (2002), and Neville and Jensen (2007) presented relational dependency networks (RDNs) that are capable of expressing and reasoning with dependencies in a relational setting, trying to overcome the acyclicity requirement problem that prevents learning arbitrary dependencies and limits the applicability of directed PRMs, such as relational Bayes networks (RBNs) (Friedman et al., 1999). Similarly to RMNs, RDNs can represent and reason with arbitrary forms of autocorrelation, also structured in a cyclic manner. Furthermore, using a pseudolikelihood estimation technique, RDNs are not limited by efficiency concerns during learning as RMNs, for which the cost of inference is prohibitively expensive.

Richardson and Domingos (2006) proposed MLNs as another extension of Markov networks for relational data, combining first-order logic and a probabilistic graphical model in a single representation. They showed how the proposed framework can easily and naturally solve collective classification tasks and that social networks are typically MLNs. Given an object `o`, its attributes can be represented in MLNs as predicates of the form `A(o, v)`, where `A` is the name of an attribute and `v` is the corresponding value. The class label `y` of an object `o` can be represented as `C(o, y)`. Classification thus corresponds to inferring the truth value of `C(o, y)` for all `o`'s and `y`'s given all known `A(o, v)`. In this specific case of collective classification, the `C(a, v)` and `C(b, v)` are not independent for all `a`'s and `b`'s given the known attribute values.

Bilgic et al. (2007) proposed a general approach interleaving object classification and link prediction in a collective algorithm. Experimental results proved that the proposed approach is preferable to running collective object classification or link prediction alone.

Another issue in network analysis is related to the evolution of the network. Indeed, online affiliation networks contain information about groups that actors have formed over time. Most collective classification algorithms take advantage only of the statistical dependencies induced by the actor-actor links. Online groups provide a clustering of the actors that is more informative than inferring groups based on actor-actor links. Zheleva et al. (2010) provided a method for classification with higher-order Markov random field models combining information from both the social network and the affiliation network.

4.2 Product recommendation

Another interesting application of SRL in social network analysis is to study the network structure with the aim of obtaining a content-based recommendation system.

Fouss et al. (2007) have proposed an approach for collaborative recommendation on the real world movie database MovieLens that naturally fits into the SRL framework.

Their work views a database as a collection of sets of elements connected by relationships. The graph structure of the database is exploited to compute, with a Markov-chain model, a similarity measure between elements. In particular, they compute quantities providing similarity measures between any pair of elements of a connected graph. These similarity measures are then used to compare items that are not necessarily directly connected.

Xu et al. (2008, 2010) discussed how the infinite hidden relational models (IHRMs) approach can be used to model and analyse social networks. In the proposed IHRM-based social network model, each edge is associated with a random variable and the probabilistic dependencies among these variables are described by the relational structure. In a hidden relational model (HRM), a hidden variable (i.e., unknown attributes) is introduced for each node of the network. Attributes of a node only depend on its hidden variable, and a relationship only depends on the hidden variables of the nodes involved in the relationship. In case of known hidden variables, both attributes and relationships can be predicted. The experimental analysis performed on the MovieLens social network studied the cooperative effect in a recommendation framework where both user properties and item properties are taken into account. The results proved that the IHRM provides good prediction accuracy for user preference on movies.

Rettinger et al. (2011) presented a method to implement and learn context-sensitive trust using SRL in the form of a Dirichlet process mixture model called infinite hidden relational trust model (IHRTM) empirically evaluated on user-ratings gathered from eBay. The main result of the proposed approach is the possibility for the truster to characterise the structure of a trust-situation providing meaningful trust assessments. IHRTM is based on IHRM (Xu et al., 2006; Kemp et al., 2006) and provides an elegant way to combine content-based predictions with collaborative-filtering predictions by exploiting regularities in the relations. The method has been tested on real world data from eBay modelled as a social trust network, proving its ability to characterise a trust-situation, its predictive performance concerning trust values, and its learning efficiency in the context of dynamic behaviour of non-stationary trustees.

Taranto et al. (2012) propose to use a framework based on probabilistic graphs that fits in SRL, in order to deal with collaborative filtering problems. In this framework, relationships among users and items and their corresponding likelihood are encoded in a probabilistic graph that can then be used to infer the probability of existence of a link between a user and an item involved in the graph. In order to solve collaborative filtering tasks the framework uses an approximate inference method adopting a constrained simple path query language. The performance of the proposed approach is reported when applied to the real world MovieLens database.

4.3 Entity resolution

Entity resolution regards the problem of determining which references in the data refer to the same underlying real world entity. The problem is caused by the merging into a single database of data from multiple databases that gives duplicate records which are not always unique identifiers of the entities thus causing ambiguity. Entity resolution has been tackled in many research areas under names such as de-duplication, data integration, co-reference resolution, object consolidation, record linkage, etc.

Traditionally, entity resolution has been solved adopting attribute similarity measures, where the co-reference degree between two entities is computed using the similarity scores between their attributes. Recent approaches try to consider structural similarity, as reported in Dong et al. (2005), and Kalashnikov and Mehrotra (2006) where inter-entities relationships and associations between references are considered: The resolution of entities of one type is helped by resolution of entities of related types. Indeed, relationships among entity references may be represented as a graph, sometimes referred as *reference graph* (Bhattacharya and Getoor, 2007), where the nodes are the entity references and edges indicate references which co-occur. These collective entity matching techniques, that use the relational information to make all the matching decisions collectively, have been shown to significantly outperform conventional approaches in terms of accuracy.

Pasula et al. (2003) studied the entity resolution problem proposing a generative relational approach based on the use of a relational probability model (RPMs) (Friedman et al., 1999) that explicitly captures the dependencies among multiple co-reference decisions. McCallum and Wellner (2004) solved the same collective problem proposing a discriminative approach that incorporates the transitive closure step into the statistical model.

Singla and Domingos (2005) proposed a method, based on conditional random fields, to solve the entity resolution problem in a collective manner. Simultaneous inferences are made for all candidate match pairs, allowing information to propagate from one candidate match to another via the attributes they have in common. The proposed model can be viewed as a form of RMN (Taskar et al., 2007).

The same authors (Singla and Domingos, 2006) presented a formulation of the entity resolution problem incorporating many non-independent and identically distributed approaches that takes advantage of SRL. They use MLNs to propose a unifying framework for entity resolution, and show, with experiments on two citation databases (Cora and BibServ), how MLNs can be used as a valuable framework to build entity resolution systems.

Bhattacharya and Getoor (2007) considered the approach of *collective* entity resolution: if two references refer to the same entity, then one may make additional inferences about their related references. They motivated entity resolution as a clustering problem and proposed a relational clustering algorithm for collective relational entity resolution. Given a similarity measure between pairs of references, entity resolution is posed as a clustering problem where the aim is to cluster the references so that only those that correspond to the same entity are assigned to the same cluster. They adopted three real world datasets describing publications in several different scientific research areas (CiteSeer, arXiv and BioBase). The goal was to use co-author relationships in the papers to help at discovering the underlying author entities in the domain and mapping the author references to the discovered author entities.

Rastogi et al. (2011) proposed a framework for scaling collective entity resolution algorithms using MLNs. The framework allows the modelling of entity resolution algorithms as black-boxes that take in a set of entities along with a collection of evidence, and output a set of matches. Running the entity resolution algorithm on the entire dataset is avoided and approximated by running multiple instances of the algorithm on several small subsets of the entities, and passing a message-set across the instances to exchange information between different runs of the matcher.

4.4 Link prediction

As already surveyed in Getoor and Diehl (2005), link prediction is one of the main task naturally modelled and solved by many SRL frameworks. The conjugation of statistical techniques, typically suitable for handling large amounts of noisy data, with relational ones, that are sufficiently powerful to deal with complex descriptions directly extracted from relational databases, is the key to tackle both difficulties of the link prediction task: the statistical component deals with the huge quantity of data, noise and real-valued features; the relational one copes with their complex representation. Additionally, a feature generation technique performs the selection of discriminant features, usually carried out manually by an expert in statistical learning.

Popescul et al. (2003) proposed to build link prediction models using structural logistic regression, an extension of logistic regression to modelling relational data that exploits aggregate operators. Logistic regression is a technique to learn discriminative models based on conditional class probabilities by tuning regression coefficients in order to maximise conditional likelihood (within a range of model complexity that avoids over-fitting). In particular, upgrading the basic logistic regression approach allows to dynamically select the relationships of interest, saving time and space. Specifically, the upgrading consists in feature generation from relational data, formulated as a top-down, breadth-first search in the space of relational database queries, represented as a directed acyclic graph ordered by some kind of generality relationship. Nodes are first-order expressions treated as database queries. Instead of simple Boolean values, they yield corresponding tables of all satisfying variable bindings, on which different aggregations can be performed using standard SQL operators, resulting in Boolean or real-valued features. A peculiar refinement operator is defined to expand search nodes to their most general specialisations. Statistical information criteria are used dynamically during the search to determine which features are to be included into the model. Clearly, this means that the features selected to build the model might not be present in future instances to be classified, differently from the classical attribute-value setting in which the set of features is pre-defined and fixed for all items (Popescul and Ungar, 2003).

As reported in Richardson and Domingos (2006), the formulation of the link prediction problem in MLNs is identical to that of collective classification, with the only difference that the goal is to infer the value of the relations between objects instead of the class of the objects.

A recent promising approach to model social networks and solve the problem of link prediction is the probabilistic logic ProbLog proposed by De Raedt et al. (2007). ProbLog attempts to plug probability handling directly into a logic reasoner, resulting in an extension of the Prolog language. In the new setting, the outcome of a query represents the probability of its success in a randomly sampled programme in which clauses are tagged with their probability of being true. This approach was inspired by biological networks, consisting of items (concepts) linked by edges labelled with mutually independent probabilities, and thus their application to social network analysis (and specifically to link prediction) is straightforward. Adopting the ProbLog language, Taranto et al. (2011a) proposed a link-based classifier that can improve the accuracy of a classical k-nearest neighbour approach when applied to image classification. From a set

of images, a probabilistic network is constructed by connecting any two images that share similar features. The probability of the edges denotes the strength of the similarity. The similarity between two images not directly connected is computed exploiting their probabilistic connections.

Zheleva et al. (2008) studied the predictive power of overlaying friendship and family ties relationships among the participants in a social network bridging approaches based on structural equivalence and community detection. In particular, the proposed approach to link prediction in multi-relational social networks is based on the use of both attribute and structural features. The main focus was to study how group membership can significantly aid in accurate link prediction.

Bilgic and Getoor (2009) proposed an active learning technique for network data while utilising links to select better examples to label. The authors extended the three classical tasks adopted in a typical utility-based active learning technique in order to utilise the links in the network.

4.5 Community detection

An interesting task in social network analysis is represented by the identification of subgroups or communities, as their discovery can be used for further analysis such as visualisation, viral marketing, determining the causal factors of group formation, detecting group evolution or stable clusters.

A community is defined as a group of actors with frequent interactions occurring among each other. The simplest interaction that can be found is local interaction. However, more interesting interactions that one would like to discover are the non-local dependencies between actors in the network. In this direction, in Xu et al. (2008, 2010), the authors investigated the possibility of applying IHRM to model and analyse social networks. To this aim, in their model, each actor is associated with a random variable and the probabilistic dependencies between such variables are specified by the model based on the relational structure. In this way, the hidden variables, one for each actor, are able to bear information that can grasp the non-local dependencies in the network. The approach was tested on the Sampson's monastery data obtaining communities that were quite close to the real groups.

Social networks are generally large and dynamic networks, where new links and contents are created every day, and where relations between actors are not clearly defined. Some works face this problem by exploiting models that consider the different types of relations between entities as links in a network. Specifically, in Bhattacharya and Getoor (2004), a bottom-up agglomerative clustering algorithm is proposed to partition links in a network into clusters. Successively such links are exploited in a relational probabilistic model in order to group entities of the network connected by the discovered links in communities. Furthermore, in Kubica et al. (2002), in addition to link evidence even attributes on entities are simultaneously considered to discover groups. The group detection algorithm uses a Bayesian network to group entities from two datasets, demographic data describing the entities and link data. In Wang et al. (2005), attributes on entities, link evidence and attributes on link evidence are exploited in a relational structure of the network to detect communities. In details, the relational structure of the network is generated using a probabilistic generative model of entity relationships and textual attributes that simultaneously discover groups among the entities and topics among the corresponding texts.

4.6 Information diffusion and role analysis

Understanding the mechanism governing how information diffuses through social networks has implications for marketing, sociology, journalism, and so on. Models of network diffusion may be used to study, for instance, product recommendation systems and viral marketing, as reported in Richardson and Domingos (2002), Domingos (2005), Leskovec et al. (2006a, 2006b) and Sharara et al. (2011).

In Domingos and Richardson (2001), Richardson and Domingos (2002), and Domingos (2005), the authors proposed to model the customer's network value. Instead of viewing a market as a set of independent entities, they view it as a social network and model it as a Markov random field. Experiments showed that the proposed model allowed to achieve much higher profits compared to ignoring interactions among customers and the corresponding network effects, as traditional marketing does. In particular, the authors try to model a *customer network value* where a customer value should represent the expected profit from sales to a customer. They model how likely each customer is to buy some product considering both the properties of the customer and product, and the *influence* (word of mouth) of his neighbours in the network.

There are two issues to be considered in information diffusion: firstly, the spread of recommendations/opinions is not standard, because their effect changes according to the trust of the receiver with respect to the sender of the messages and to their having the same tastes and needs; then, partly because of this and partly because of the inherently evolving nature of the network elements (typically humans), the network tends to be dynamic, which might have a significant impact on the information spread. While most traditional works in the literature on diffusion modelling have underestimated either or both these questions, Sharara et al. (2011) specifically focused on them, and proposed the differential adaptive diffusion model as a solution. The network is represented, as usual, as a graph whose nodes are individuals and whose links, representing social relationships, are weighted with the confidence in the corresponding recommendations; additionally, a function determines the preference of a user for a product. Then, selected nodes are chosen to start the diffusion process considering as if they adopted the given product. These nodes may activate neighbour nodes according to any traditional diffusion model, and those neighbours that actually adopt the product according to such a model in turn may cause the spread of adoption. After the spread has come to an end, a kernel function is exploited on the network to re-weight link confidence according to the actual adoptions that took place.

Another question in the same direction is how to identify the role of network elements with respect to diffusion. For instance, it might be interesting to identify opinion leaders having a significant influence on the adoption of a given product with their recommendations. Solving manually the problem using primary sources (surveys and interviews) is costly and difficult, hence the interest for automatic techniques that work on the data provided by observation of the network. Unfortunately, simple notions of centrality based on the number of connections of an element are not sufficiently predictive, requiring more sophisticated techniques. Sharara et al. (2010) proposed to solve the problem using an active learning framework, in which the learner autonomously gathers information (examples) for refining the model. In particular, the network elements on which applying primary sources are selected as the minimum set of respondents needed to classify a given percentage of opinion leaders in the network (a technique called active survey) based on secondary sources (i.e., information expressed

by the networks in which the elements participate). Such sources consist of both node features and edge features, exploited in a probabilistic inference based on a likelihood function of a new candidate leader being nominated by a respondent to a primary source. The current set of leaders is updated according to this choice, and another loop is run, until the desired percentage of leaders has been found. The initial seed for this iterative procedure is selected by identifying communities in the network, and then selecting a representative from each community.

Delaney et al. (2010) applied SRL algorithms to predict leadership roles of individuals in a group based on patterns of activity, communication, and individual attributes. The authors focused on data collection on criminal and terror networks whose straightforward use includes manual analysis of groups and individuals involved in nefarious activity to inform key decision makers tasked with preventing future bombings or other violent attacks.

5 Conclusions

In addition to their computational and data management capabilities, computers are nowadays mostly appreciated because of their being a means for communication and information sharing among people all over the world. The spread of the internet has allowed/caused several complex interactions to take place among people, which resulted in the birth of huge information networks based on these interactions. Social networks are determined by various kinds of social relationships that connect people, and potentially represent an invaluable source of information that can be exploited for scientific and commercial purposes. SNM is the corresponding research area, aimed at extracting information about the network objects and behaviour that cannot be obtained based on the explicit/implicit description of the objects alone, ignoring their explicit/implicit relationships. Differently from other kinds of data on which information extraction tasks have been carried out in the past, social networks are characterised by a huge size and by the inherently relational setting as distinguishing peculiarities. As a consequence, their mining requires new techniques to gather this information.

SRL provides very promising approaches to deal with SNM, since it combines expressive representation formalisms, able to model complex relational networks, with statistical methods able to handle uncertainty about objects and relations. The capability of SRL models to naturally deal with relational representations, such as complex networks, represents its important characteristic in providing algorithms and methods able to outperform traditional propositional based techniques. This paper proposed a survey of some SRL formalisms and techniques adopted to solve some SNM tasks.

References

Acar, E., Dunlavy, D.M. and Kolda, T.G. (2009) 'Link prediction on evolving data using matrix and tensor factorizations', in *Proceedings of the IEEE International Conference on Data Mining Workshops*, pp.262–269, IEEE Computer Society.

Aggarwal, C.C., Zhao, Y. and Yu, P.S. (2011) 'Outlier detection in graph streams', in *Proceedings of the 27th International Conference on Data Engineering*, pp.399–409, IEEE Computer Society.

Akoglu, L., McGlohon, M. and Faloutsos, C. (2009) 'Anomaly detection in large graphs', in CMU-CS-09-173 Technical Report.

Backstrom, L. and Leskovec, J. (2011) 'Supervised random walks: predicting and recommending links in social networks', in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pp.635–644, ACM.

Bhattacharya, I. and Getoor, L. (2004) 'Deduplication and group detection using links', in *ACM SIGKDD Workshop on Link Analysis and Group Detection (LinkKDD)*.

Bhattacharya, I. and Getoor, L. (2007) 'Collective entity resolution in relational data', *ACM Transactions on Knowledge Discovery from Data*, March, Vol. 1, No. 1, pp.1–36.

Bilgic, M. and Getoor, L. (2009) 'Link-based active learning', in *NIPS Workshop on Analyzing Networks and Learning with Graphs*.

Bilgic, M., Namata, G.M. and Getoor, L. (2007) 'Combining collective classification and link prediction', in *Proceedings of the 7th IEEE International Conference on Data Mining Workshops*, pp.381–386, IEEE Computer Society.

Buskens, V. and Yamaguchi, K. (1999) 'A new model for information diffusion in heterogeneous social networks', *Sociological Methodology*, Vol. 29, No. 1, pp.281–325.

Cao, B., Liu, N.N. and Yang, Q. (2010) 'Transfer learning for collective link prediction in multiple heterogenous domains', in Fürnkranz, J. and Joachims, T. (Eds.): *Proceedings of the 27th International Conference on Machine Learning*, pp.159–166, Omnipress.

Chakrabarti, D. (2004) 'Autopart: parameter-free graph partitioning and outlier detection', in *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp.112–124, Springer-Verlag New York, Inc.

Chakrabarti, S., Dom, B. and Indyk, P. (1998) 'Enhanced hypertext categorization using hyperlinks',iIn *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pp.307–318, ACM.

Costa, V.S. and Cussens, J. (2003) 'CLP(BN): constraint logic programming for probabilistic knowledge', in *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, pp.517–524, Morgan Kaufmann.

Cussens, J. (2007) 'Logic-based formalisms for statistical relational learning', in Getoor, L. and Taskar, B. (Eds.): *Introduction to Statistical Relational Learning*, pp.269–290, MIT Press.

David Heckerman, D.K. and Meek, C. (2007) 'Probabilistic entity-relationship models, PRMs, and plate models', in Getoor, L. and Taskar, B. (Eds.): *Introduction to Statistical Relational Learning*, pp.201–239, MIT Press.

De Raedt, L., Demoen, B., Fierens, D., Gutmann, B., Janssens, G., Kimmig, A., Landwehr, N., Mantadelis, T., Meert, W., Rocha, R., Santos Costa, V., Thon, I. and Vennekens, J. (2008a) 'Towards digesting the alphabet-soup of statistical relational learning', in Roy, D., Winn, J., McAllester, D., Mansinghka, V. and Tenenbaum, J. (Eds.): *Proceedings of the 1st Workshop on Probabilistic Programming: Universal Languages, Systems and Applications*.

De Raedt, L., Frasconi, P., Kersting, K. and Muggleton, S. (Eds.) (2008b) 'Probabilistic inductive logic programming', *LNCS*, Vol. 4911, Springer.

De Raedt, L., Kimmig, A. and Toivonen, H. (2007) 'ProbLog: a probabilistic prolog and its application in link discovery', in *Proceedings of 20th International Joint Conference on Artificial Intelligence*, pp.2468–2473, AAAI Press.

de Salvo Braz, R., Amir, E. and Roth, D. (2005) 'Lifted first-order probabilistic inference', in Kaelbling, L.P. and Saffiotti, A. (Eds.): *Nineteenth International Joint Conference on Artificial Intelligence*, pp.1319–1325.

Delaney, B., Fast, A.S., Campbell, W.M., Weinstein, C.J. and Jensen, D.D. (2010) 'The application of statistical relational learning to a database of criminal and terrorist activity', in *Proceedings of the SIAM International Conference on Data Mining*, pp.409–417.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, No. 1, pp.1–38.

Di Mauro, N., Basile, T.M., Ferilli, S. and Esposito, F. (2011) 'Optimizing probabilistic models for relational sequence learning', in Kryszkiewicz, M., Rybinski, H., Skowron, A. and Ras, Z.W. (Eds.): *19th International Symposium on Methodologies for Intelligent Systems, LNCS*, pp.240–249, Springer.

Domingos, P. (2005) 'Mining social networks for viral marketing', *IEEE Intelligent Systems*, Vol. 20, No. 1, pp.80–82.

Domingos, P. and Richardson, M. (2001) 'Mining the network value of customers', in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.57–66, ACM.

Dong, X., Halevy, A. and Madhavan, J. (2005) 'Reference reconciliation in complex information spaces', in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, pp.85–96, ACM.

Fouss, F., Pirotte, A., Renders, J-M. and Saerens., M. (2007) 'Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation', *IEEE Transaction on Knowledge and Data Engineering*, Vol. 19, No. 3, pp.355–369.

Friedman, N., Getoor, L., Koller, D. and Pfeffer, A. (1999) 'Learning probabilistic relational models', in Dean, T. (Ed.): *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pp.1300–1309, Morgan Kaufmann.

Gao, J., Liang, F., Fan, W., Wang, C., Sun, Y. and Han, J. (2010) 'On community outliers and their efficient detection in information networks', in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.813–822, ACM.

Gargi, U., Lu, W., Mirrokni, V.S. and Yoon, S. (2011) 'Large-scale community detection on youtube for topic discovery and exploration', in Adamic, L.A., Baeza-Yates, R.A. and Counts, S. (Eds.): *Proceedings of the 5th International Conference on Weblogs and Social Media*, The AAAI Press.

Getoor, L. (2003) 'Link mining: a new data mining challenge', *SIGKDD Explorations Newsletter*, July, Vol. 5, pp.84–89.

Getoor, L. and Diehl, C.P. (2005) 'Link mining: a survey', *SIGKDD Explorations Newsletter*, Vol. 7, pp.3–12.

Getoor, L. and Taskar, B. (Eds.) (2007) *Introduction to Statistical Relational Learning*, MIT Press, Cambridge, Massachusetts.

Gleave, E., Welser, H.T., Lento, T.M. and Smith, M.A. (2009) 'A conceptual and operational definition of 'social role' in online community', in *Proceedings of the 42nd Hawaii International Conference on System Sciences*, pp.1–11, IEEE Computer Society.

Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A. (2004) 'Information diffusion through blogspace', in *Proceedings of the 13th International Conference on World Wide Web*, pp.491–501, ACM.

Hasan, M.A., Chaoji, V., Salem, S. and Zaki, M. (2006) 'Link prediction using supervised learning', in *Proc. of SDM 06 Workshop on Link Analysis, Counter Terrorism and Security*.

Hohwald, H., Cebrian, M., Canales, A., Lara, R. and Oliver, N. (2009) 'Inferring unobservable inter-community links in large social networks', in *Proceedings of the 2009 International Conference on Computational Science and Engineering*, Vol. 4, pp.375–380, IEEE Computer Society.

Hsu, T.T-s., Liau, C-J. and Wang, D-W. (2008) 'Logical definability in social position analysis', in *Proceeding of International Conference on Granular Computing*, pp.35–38, IEEE.

Hui, P., Yoneki, E., Chan, S.Y. and Crowcroft, J. (2007) 'Distributed community detection in delay tolerant networks', in *Proceedings of 2nd ACM/IEEE International Workshop on Mobility in the Evolving Internet Architecture*, pp.7:1–7:8, ACM.

Jaeger, M. (1997) 'Relational Bayesian networks', in Geiger, D. and Shenoy, P.P. (Eds.): *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*, pp.266–273, Morgan Kaufmann.

Jensen, D., Neville, J. and Gallagher, B. (2004) 'Why collective inference improves relational classification', in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.593–598, ACM.

Kalashnikov, D.V. and Mehrotra, S. (2006) 'Domain-independent data cleaning via analysis of entity-relationship graph', *ACM Trans. Database Syst*., Vol. 31, No. 2, pp.716–767.

Karamon, J., Matsuo, Y., Yamamoto, H. and Ishizuka, M. (2007) 'Generating social network features for link-based classification', in *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp.127–139, Springer-Verlag, Berlin, Heidelberg, ISBN: 978-3-540-74975-2.

Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T. and Ueda, N. (2006) 'Learning systems of concepts with an infinite relational model', in *Proceedings of the 21st National Conference on Artificial Intelligence*, Vol. 1, pp.381–388, AAAI Press.

Kempe, D., Kleinberg, J. and Tardos, E. (2003) 'Maximizing the spread of influence through a social network', in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.137–146, ACM.

Kersting, K. and De Raedt, L. (2007) 'Bayesian logic programming: theory and tool', in Getoor, L. and Taskar, B. (Eds.): *Introduction to Statistical Relational Learning*, Chapter 10, MIT Press.

Kersting, K., De Raedt, L. and Kramer, S. (2000) 'Interpreting Bayesian logic programs', in *Proceedings of the Work-in-Progress Track at the 10th International Conference on Inductive Logic Programming*, pp.138–155.

Kubica, J., Moore, A., Schneider, J. and Yang, Y. (2002) 'Stochastic link and group detection', in *18th National Conference on Artificial Intelligence*, pp.798–804, American Association for Artificial Intelligence.

Lafferty, J. and Lebanon, G. (2002) 'Information diffusion kernels', in *Advances in Neural Information Processing Systems*, Vol. 15, pp.375–382, MIT Press.

Landwehr, N., Kersting, K. and De Raedt, L. (2005) 'nFOIL: integrating naïve Bayes and FOIL', in Veloso, M.M. and Kambhampati, S. (Eds.): *Proceedings of the 20th National Conference on Artificial Intelligence and the 17th Innovative Applications of Artificial Intelligence Conference*, pp.795–800, AAAI Press/The MIT Press.

Landwehr, N., Passerini, A., De Raedt, L. and Frasconi, P. (2006) 'kFOIL: learning simple relational kernels', in *Proceedings, The 21st National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, AAAI Press.

Leroy, V., Cambazoglu, B.B. and Bonchi, F. (2010) 'Cold start link prediction', in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.393–402. ACM.

Leskovec, J., Adamic, L.A. and Huberman, B.A. (2006a) 'The dynamics of viral marketing', in *Proceedings of the 7th ACM Conference on Electronic Commerce*, pp.228–237, ACM.

Leskovec, J., Singh, A. and Kleinberg, J.M. (2006b) 'Patterns of influence in a recommendation network', in Ng, W.K., Kitsuregawa, M., Li, J. and Chang, K. (Eds.): *Proceeding of the 10th Pacific Asia Conference on Advances in Knowledge Discovery and Data Mining, LNCS*, Vol. 3918, pp.380–389, Springer.

Leskovec, J., Lang, K.J. and Mahoney, M. (2010) 'Empirical comparison of algorithms for network community detection', in *Proceedings of the 19th International Conference on World Wide Web*, pp.631–640, ACM.

Leung, I.X.Y., Hui, P., Lio, P. and Crowcroft, J. (2009) 'Towards real-time community detection in large networks', *Physical Review E*, Vol. 79, No. 6, pp.66–107.

Liben-Nowell, D. and Kleinberg, J. (2007) 'The link-prediction problem for social networks', *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 7, pp.1019–1031.

Lichtenwalter, R.N. and Chawla, N.V. (2011) 'Lpmade: link prediction made easy', *Journal of Machine Learning Research*, August, Vol. 12, pp.2489–2492.

Lichtenwalter, R.N., Lussier, J.T. and Chawla, N.V. (2010) 'New perspectives and methods in link prediction', in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.243–252, ACM.

Liu, W. and Lu, L. (2006) 'Link prediction based on local random walk', *EPL*, Vol. 89, p.58007.

Lozano, S., Duch, J. and Arenas, A. (2006) 'Community detection in a large social dataset of European projects', in *Proceedings of the 6th SIAM International Conference on Data Mining*.

Macskassy, S.A. and Provost, F. (2007) 'Classification in networked data: a toolkit and a univariate case study', *Journal of Machine Learning Research*, May, Vol. 8, pp.935–983.

Maiya, A.S. and Berger-Wolf, T.Y. (2010) 'Sampling community structure', in *Proceedings of the 19th International Conference on World Wide Web*, pp.701–710, ACM.

McCallum, A. and Wellner, B. (2004) 'Conditional models of identity uncertainty with application to noun co-reference', in *Proceedings of Advances in Neural Information Processing Systems 17*.

McCallum, A., Corrada-emmanuel, A. and Wang, X. (2005) 'Topic and role discovery in social networks', in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp.786–791.

Meo, P.D., Ferrara, E., Fiumara, G. and Provetti, A. (2011) 'Generalized Louvain method for community detection in large networks', in *11th International Conference on Intelligent Systems Design and Applications*, pp.88–93.

Miller, K., Griffiths, T. and Jordan, M. (2009) 'Nonparametric latent feature models for link prediction', in Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C.K.I. and Culotta, A. (Eds.): *Advances in Neural Information Processing Systems*, Vol. 22, pp.1276–1284.

Muggleton, S. (1991) 'Inductive logic programming', *New Generation Computing*, Vol. 8, No. 4, pp.295–318.

Muggleton, S. (1996) 'Stochastic logic programs', in De Raedt, L. (Ed.): *Advances in Inductive Logic Programming*, pp.254–264, IOS Press.

Neville, J. and Jensen, D. (2007) 'Relational dependency networks', *Journal of Machine Learning Research*, May, Vol. 8, pp.653–692.

Neville, J., Jensen, D., Friedland, L. and Hay, M. (2003) 'Learning relational probability trees', in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.625–630.

Ng, R. and Subrahmanian, V.S. (1992) 'Probabilistic logic programming', *Journal Information and Computation*, Vol. 101, No. 2, pp.150–201.

Ouimet, M., Landry, R. and Amara, N. (2004) 'Network positions and radical innovation: a social network analysis of the Quebec optics/photonics cluster', in *Proceedings of the DRUID Summer Conference 2004: Industrial Dynamics, Innovation and Development*, pp.786–791.

Pasula, H., Marthi, B., Milch, B., Russell, S. and Shpitser, I. (2003) 'Identity uncertainty and citation matching', in *Advances in Neural Information Processing*, Vol. 16, pp.1401–1408, MIT Press.

Pearl, J. (1991) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Francisco, CA, USA.

Poole, D. (1993) 'Probabilistic horn abduction and Bayesian networks', *Artificial Intelligence*, Vol. 64, No. 1, pp.81–129.

Poole, D. (2003) 'First-order probabilistic inference', in *8th International Joint Conference on Artificial Intelligence*, pp.985–991.

Popescul, A. and Ungar, L.H. (2003) 'Statistical relational learning for link prediction', in *IJCAI Workshop on Learning Statistical Models from Relational Data*.

Popescul, A., Popescul, R. and Ungar, L.H. (2003) 'Structural logistic regression for link analysis', in *Proceedings of the 2nd International Workshop on Multi-Relational Data Mining*, pp.92–106, ACM Press.

Rastogi, V., Dalvi, N. and Garofalakis, M. (2011) 'Large-scale collective entity matching', in *Proceedings of Very Large Database Endowment*, Vol. 4, pp.208–218, VLDB Endowment.

Rettinger, A., Nickles, M. and Tresp, V. (2011) 'Statistical relational learning of trust', *Machine Learning*, Vol. 82, No. 2, pp.191–209.

Richardson, M. and Domingos, P. (2002) 'Mining knowledge-sharing sites for viral marketing', in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.61–70, ACM.

Richardson, M. and Domingos, P. (2006) 'Markov logic networks', *Machine Learning*, Vol. 62, Nos. 1–2, pp.107–136.

Riguzzi, F. and Di Mauro, N. (2012) 'Applying the information bottleneck to statistical relational learning', *Machine Learning Journal*, Vol. 86, No. 1, pp.89–114.

Romero, D.M., Meeder, B. and Kleinberg, J. (2011) 'Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter', in *Proceedings of the 20th International Conference on World Wide Web*, pp.695–704, ACM.

Sarkar, P., Chakrabarti, D. and Moore, A.W. (2011) 'Theoretical justification of popular link prediction heuristics', in *Proceedings of International Joint Conference on Artificial Intelligence*, pp.2722–2727.

Sato, T. and Kameya, Y. (1997) 'PRISM: a language for symbolic-statistical modeling', in *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pp.1330–1335.

Sen, P., Namata, G.M., Bilgic, M., Getoor, L., Gallagher, B. and Eliassi-Rad, T. (2008) 'Collective classification in network data', *AI Magazine*, Vol. 29, No. 3, pp.93–106.

Senator, T.E. (2005) 'Link mining applications: progress and challenges', *SIGKDD Exploration Newsletter*, December, Vol. 7, No. 2, pp.76–83, ISSN: 1931-0145.

Sharara, H., Getoor, L. and Norton, M. (2010) 'An active learning approach for identifying key opinion leaders', in *The 2nd Workshop on Information in Networks (WIN)*.

Sharara, H., Rand, W. and Getoor, L. (2011) 'Differential adaptive diffusion: understanding diversity and learning whom to trust in viral marketing', in *The 5th International AAAI Conference on Weblogs and Social Media*.

Singla, P. and Domingos, P. (2005) 'Object identification with attribute-mediated dependences', in Jorge, A., Torgo, L., Brazdil, P., Camacho, R. and Gama, J. (Eds.): *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp.297–308.

Singla, P. and Domingos, P. (2006) 'Entity resolution with Markov logic', in *Proceedings of the Sixth International Conference on Data Mining*, pp.572–582, IEEE Computer Society.

Tang, L. and Liu, H. (2010) 'Graph mining applications to social network analysis', in Aggarwal, C. and Wang, H. (Eds.): *Managing and Mining Graph Data*, pp.487–516, Springer.

Tang, L. and Liu, H. (2011) 'Leveraging social media networks for classification', *Data Mining and Knowledge Discovery*, Vol. 23, No. 3, pp.1–32.

Taranto, C., Di Mauro, N. and Esposito, F. (2011a) 'Probabilistic inference over image networks', in Agosti, M., Esposito, F., Meghini, C. and Orio, N. (Eds.): *7th Italian Research Conference on Digital Libraries and Archives, CCIS*, pp.1–13, Springer.

Taranto, C., Di Mauro, N. and Esposito, F. (2011b) 'rslda: a Bayesian hierarchical model for relational learning', in Zhang, J. and Livraga, G. (Eds.): *International Conference on Data and Knowledge Engineering*, pp.68–74, IEEE.

Taranto, C., Di Mauro, N. and Esposito, F. (2012) 'Uncertain graphs meet collaborative filtering', in Amati, G., Carpineto, C. and Semeraro, G. (Eds.): *Proceedings of the 3rd Italian Information Retrieval Workshop*, Vol. 835, pp.89–100, CEUR-WS.

Taskar, B., Abbeel, P. and Koller, D. (2002) 'Discriminative probabilistic models for relational data', in Darwiche, A. and Friedman, N. (Eds.): *Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence*, pp.485–492, Morgan Kaufmann.

Taskar, B., Abbeel, P., Wong, M. and Koller, D. (2007) 'Relational Markov networks', in Getoor, L. and Taskar, B. (Eds.): *Introduction to Statistical Relational Learning*, MIT Press.

Taskar, B., Segal, E. and Koller, D. (2001) 'Probabilistic classification and clustering in relational data', in *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Vol. 2, pp.870–876, Morgan Kaufmann Publishers Inc.

Vennekens, J., Verbaeten, S. and Bruynooghe, M. (2004) 'Logic programs with annotated disjunctions', in *Proceedings of 20th International Conference on Logic Programming*, pp.431–445, Springer.

Wang, X., Mohanty, N. and Mccallum, A. (2005) 'Group and topic discovery from relations and text', in *Proceedings of the 3rd International Workshop on Link Discovery*, ACM.

Xu, Z., Kersting, K. and Tresp, V. (2009) 'Multi-relational learning with Gaussian processes', in *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pp.1309–1314, Morgan Kaufmann Publishers Inc.

Xu, Z., Tresp, V., Rettinger, A. and Kersting, K. (2010) 'Social network mining with nonparametric relational models', in Giles, L., Smith, M., Yen, J. and Zhang, H. (Eds.): *Advances in Social Network Mining and Analysis, LNCS*, Vol. 5498, pp.77–96, Springer.

Xu, Z., Tresp, V., Yu, K. and Kriegel, H-P. (2006) 'Infinite hidden relational models', in *Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence*.

Xu, Z., Tresp, V., Yu, S. and Yu, K. (2008) 'Nonparametric relational learning for social network analysis', in *Proceedings of the 2nd ACM Workshop on Social Network Mining and Analysis*.

Yang, J. and Leskovec, J. (2010) 'Modeling information diffusion in implicit networks', in *Proceedings of the 2010 IEEE International Conference on Data Mining*, pp.599–608, IEEE Computer Society.

Zheleva, E., Getoor, L. and Sarawagi, S. (2010) 'Higher-order graphical models for classification in social and affiliation networks', in *NIPS Workshop on Networks Across Disciplines: Theory and Applications*.

Zheleva, E., Getoor, L., Golbeck, J. and Kuter, U. (2008) 'Using friendship ties and family circles for link prediction', in *2nd ACM SIGKDD Workshop on Social Network Mining and Analysis*.