entities using existing web technologies, which useful data to provide when a URI is looked up, how to model data, and what are suitable and best ways to interlink and cross-reference data sets on the web Heath and Bizer (2011).

LOD has seen a tremendous growth in the last years. The resulting distributed graph of interlinked entities on the web is commonly referred to as the LOD cloud (see Fig. 1) and spans hundreds of data sources providing more than 30 billion RDF triples (Weikum 2013). Thereby, LOD covers various domains ranging from media-related content, social networks and user-generated contents, bibliographic data, life sciences, medicine, biology, geographic data, to governmental data. Furthermore, some data sources such as DBPedia and YAGO provide general, cross-domain information and thereby play a pivotal role in connecting also data from very different domains.

Cross-References

▸ RDF

References

Berners-Lee T (2007) Linked Data. http://www.w3.org/DesignIssues/LinkedData.html. Accessed 19 Aug 2013

Cyganiak R, Jentzsch A (2011) Linking Open Data cloud diagram. http://lod-cloud.net/

Heath T, Bizer C (2011) Linked Data: evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology, vol 1(1), 1st edn. Morgan & Claypool, San Rafael, pp 1–136

Weikum G (2013) Where's the data in the big data wave? http://wp.sigmod.org/?p=786. Accessed 19 Aug 2013

Recommended Reading

Heath T, Bizer C (2011) Linked Data: evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology, vol 1(1), 1st edn. Morgan & Claypool, San Rafael, pp 1–136

Link: Edge

▸ Community Detection, Current and Future Research Trends
▸ Networks in the Twenty-First Century

Link Prediction

▸ Imputation of Missing Network Data: Some Simple Procedures
▸ Inferring Social Ties

Link Prediction: A Primer

Nitesh V. Chawla[1] and Yang Yang[2]
[1]Department of Computer Science and Engineering, Interdisciplinary Center for Network Science and Applications (iCeNSA), University of Notre Dame, Notre Dame, IN, USA
[2]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA

Synonyms

Edge prediction; Relationship extraction; Social ties inferring

Glossary

Common Neighbors For two nodes, u and v, the set of their common neighbors is defined as $N = \Gamma(u) \cap \Gamma(v)$, and correspondingly the size of set N is $|\Gamma(u) \cap \Gamma(v)|$ (Newman 2001)

Jaccard Coefficient *Jaccard coefficient* (Liben-Nowell and Kleinberg 2007) is a normalization of the *common neighbors* metric, which is defined as $JC(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$

Adamic/Adar The *Adamic/Adar* (Adamic Lada and Adar 2003) metric is defined as $AA(u,v) = \sum_{n \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log|\Gamma(n)|}$, where $n \in N, N = \Gamma(u) \cap \Gamma(v)$ is the set of common neighbors of u and v

Preferential Attachment The *preferential attachment* (Barabasi et al. 2002) metric is the multiplication of nodes u and v's degrees, $PA(u,v) = \Gamma(u).\Gamma(v)$

Katz Leo Katz proposed this metric in Katz (1953); *Katz* metric sums all paths that exist between nodes u and v and penalizes the contribution of long paths exponentially by a factor of β^l, where l is the length of the path. The equation of *Katz* is $katz(u,v) = \sum_{l=1}^{+\infty} \beta^l . |\text{paths}_{u,v}^l|$

Graph Distance The shortest path length between two given nodes u and v

Class Imbalance In the link prediction problem, the class imbalance refers to the inherent disproportion of links that can form to links that do form

ROC A receiver operating characteristic (ROC) represents the performance trade-offs between true positives and false positives by thresholding at different decision boundaries for relative true positive rate and false positive rates

AUROC Area under receiver operating characteristic (ROC) curve (Clause et al. 2008)

Definition: The Link Prediction Problem

Link prediction, that is, predicting the formation of links in a network in the future or predicting the missing links in a network, is an active topic of research. Generally, the link prediction problem can be classified into two categories: (1) predict the likelihood of a future link or missing link between two nodes, knowing that there is no link between them in the current network, and (2) predict whether there will be a future interaction between two nodes that are observed to have an association before. The specific problem discussed in this entry falls into the latter category. Formally, the link prediction problem can be formulated as below:

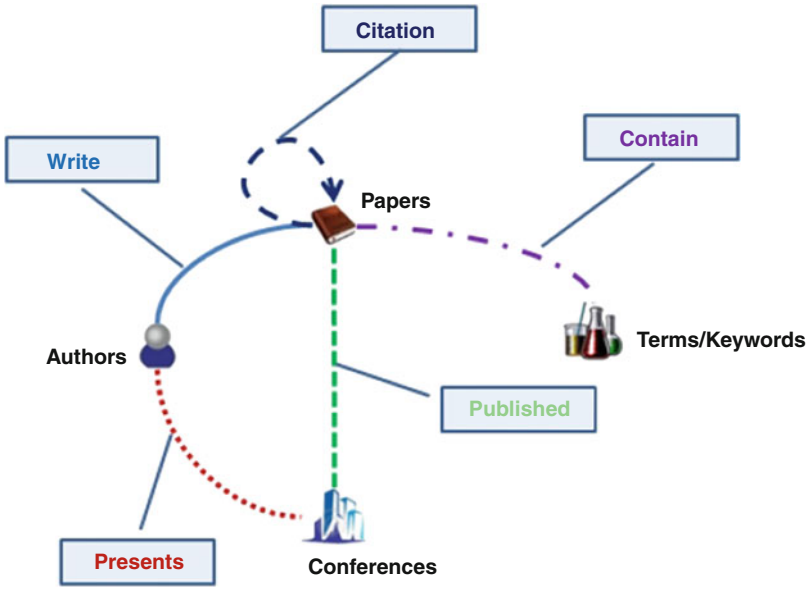
Definition 1 (Link Prediction) In a network G=(V, E), the link prediction task in such network is to predict whether there will be a link between a pair of nodes u and v, where $u, v \in V$ and $e(u,v) \notin E$.

In real world there are two kinds of networks:

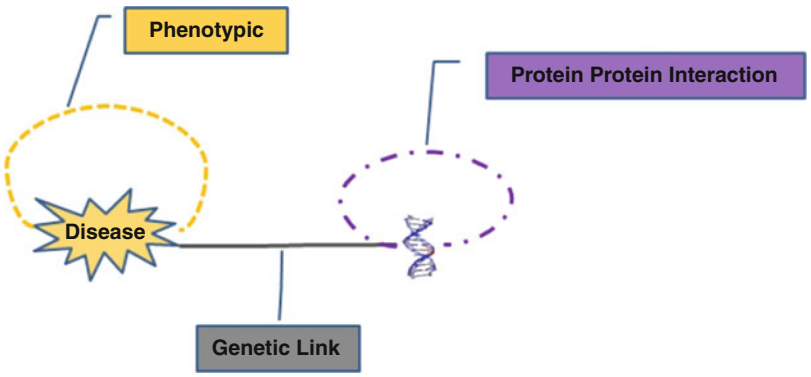
Definition 2 (Heterogeneous Network and Homogeneous Network) A network is defined as a graph $G = (V_1 \cup V_2 \ldots \cup V_I, E_1 \cup E_2 \ldots \cup E_J)$, where $V_i (i \in \{1, 2, \ldots, I\})$ represents the set of nodes of the same object type i and $E_j (j \in \{1, 2, \ldots, J\})$ represents the set of links with the same type j. When the types of nodes $I > 1$ or the types of links $J > 1$, the network is called a **heterogeneous network**; otherwise, it is a **homogeneous network**.

Introduction

Link prediction is an important task in network analysis, benefiting researchers and organizations in a variety of fields. There are a variety of techniques for link prediction, ranging from feature-based classification to probabilistic models and matrix factorization. In this entry, we mainly discuss how to solve the link prediction problem as a *supervised classification* task. Generally, we can categorize link prediction methods into two classes: (1) unsupervised methods, which extract features, such as *common neighbors*, to directly estimate the likelihood of the link, and (2) supervised methods, which train a binary classification model by extracting a set of features. Most of these methods are designed for homogeneous networks; however, many important real-world networks are heterogeneous (Viswanath et al. 2009), such as DBLP bibliographic networks and human disease-gene networks (Figs. 1 and 2) (Deng et al. 2011). The complexity of structural dependency and heterogeneity of links produces obstacles for link prediction in heterogeneous networks; however, it also provides us abundant information to learn from. Well-known topological features designed for homogeneous networks are difficult to apply to heterogeneous networks, such as *common neighbors* and *Adamic/Adar*;



Link Prediction: A Primer, Fig. 1 DBLP bibliographic network



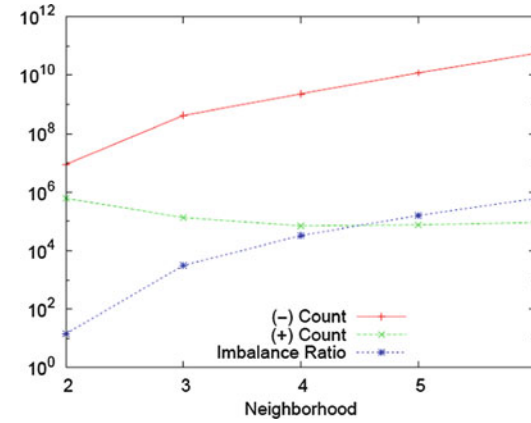
Link Prediction: A Primer, Fig. 2 Disease-gene network

inappropriate solutions lead to a loss of information. A few studies have worked on the link prediction in heterogeneous networks, from the early work (Taskar et al. 2003) to the recent work of Sun et al. (2011), Davis et al. (2011), Lichtenwalter and Chawla (2012), Yang et al. (2012), and Dong et al. (2012).

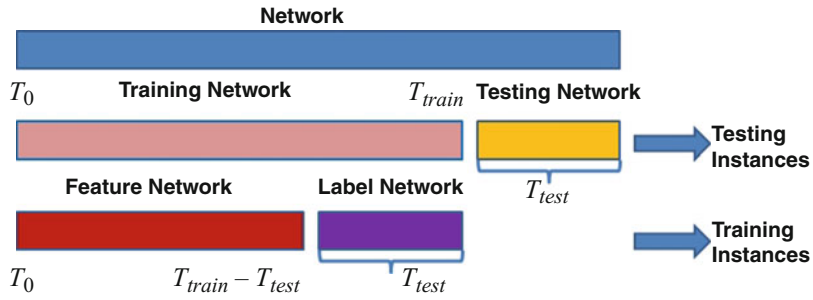
Key Points

In this entry, we mainly discuss the solutions that model the link prediction problem as a *supervised classification* task. Certain key steps are required and determine the final prediction performance (Fig. 5):

- Select effective features that describe the link likelihood appropriately. Most of the features are extracted from the network topology; sometimes non-topological features are also included into the set of features, such as location information (Scellato et al. 2011).
- Develop methods of solving class imbalance issue. For example, in Fig. 3, we can see the proportion of (−) count (potential links that do not form) to (+) count (potential links that do form) ranges from 10 to 10^6



Link Prediction: A Primer, Fig. 3 Class imbalance (cell phone network) (Lichtenwalter et al. 2010)



Link Prediction: A Primer, Fig. 4 Experimental setup

in different neighborhoods, which means the ratio between positive instances and negative instances in the classification problem is inherently disproportional.

- Appropriate experimental setup for the extraction of training set and testing set. One of the commonly used configurations is presented in Fig. 4; we use the link information from the training interval (purple area) to make predictions of future links in the test interval (yellow area).

Historical Background

Unsupervised Solutions of the Link Prediction Problem

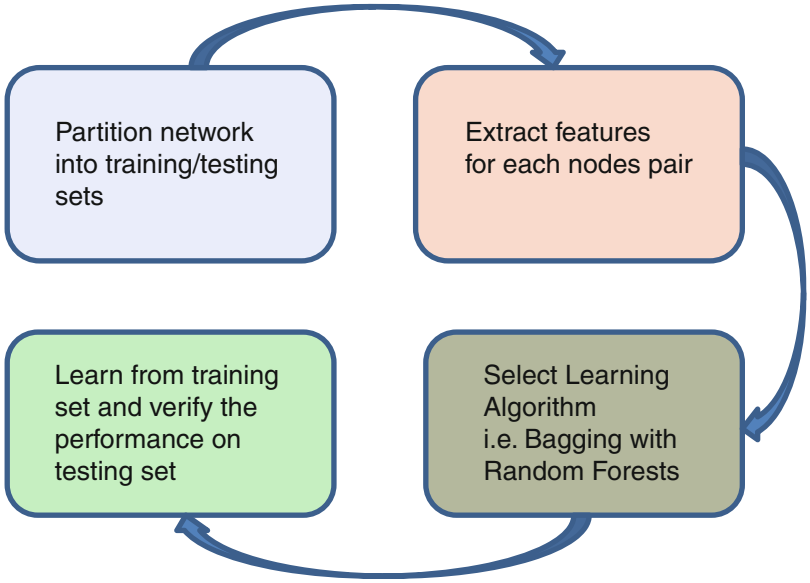
Liben-Nowell and Kleinberg (2007) proposed one of the earliest link prediction methods in social networks. Their approach is based on measures for analyzing the "proximity" of nodes in a network, including *common neighbors*, *preferential attachment*, *Adamic/Adar*, *Jaccard coefficient*, *Katz*, and *graph distance*. In the work of Liben-Nowell and Kleinberg (2007), they proposed there could be better solutions by using machine learning technologies.

Link Prediction as a Supervised Task

As we have discussed, the techniques for the link prediction task range from feature-based classification to probabilistic models and matrix factorization:

Probabilistic Models for Link Prediction: There is a stream of research that focus on modeling the posterior probability of the co-occurrence of the node pairs, such as *Markov random field*-based probabilistic model proposed by Wang et al. (2007) and Kashima's work Kashima et al. (2006) on network evolution-based probabilistic model.

Matrix Factorization for Link Prediction: In the work of (Menon and Elkan 2011; Kunegis and Lommatzsch 2009), the observed graph



Link Prediction: A Primer, Fig. 5 Supervised learning process

is represented as a matrix $\{0, 1, ?\}^{n \times n}$, where 0 denotes a known absent link, 1 denotes a known present link, and ? denotes an unknown status link. The goal is to make predictions for the ? entries by using the *matrix factorization* technique.

Feature-Based Classification for Link Prediction: The link prediction problem can be modeled as a *supervised classification* task (Murata and Moriyasu 2007; Lichtenwalter et al. 2010; Scellato et al. 2011; Sun et al. 2012; Qiu et al. 2011), where each instance corresponds to a pair of nodes in the network. One of the representative works that model the link prediction task as *classification* problem is a general supervised framework called HPLP developed by Lichtenwalter et al. (2010) (Fig. 5). By using the ability of supervised frameworks and oversampling strategies, the HPLP framework was demonstrated to outperform most unsupervised link predictors.

Link Prediction in Social Networks

In the above section we have introduced main streams of the link prediction techniques, which work in *homogeneous* networks. However, many real-world networks are heterogeneous (Figs. 1 and 2), which motivates the research of link prediction in heterogeneous networks. Most topological features for the link prediction in homogeneous networks are difficult to apply in heterogeneous networks. The key idea of a successful heterogeneous link prediction framework is *how to capture the interrelation between different types of links and employ this information effectively and precisely*. In the following sections, we will discuss the most recent solutions of the link prediction task in heterogeneous networks.

Meta Path: Sun et al. (2011) proposed the concept of **meta path**, which is a path defined on the network schema, where nodes are object types and edges are relations between object types. Table 1 lists examples of several meta paths of length 3 between authors in the DBLP network (Fig. 1). In Table 1, two types of *meta paths* between node pairs have different *p value* in describing their latent links. The *meta path*-based features are systematically extracted from the network and used in a supervised model to predict coauthor relationship in a complex DBLP network.

MRLP: Davis et al. (2011) proposed a probabilistic method (**MRLP** (multi-relational link prediction)) that performs a triad census of two nodes in heterogeneous networks, which conveniently translates to a nonarbitrary, data-justified weighting scheme (Fig. 6).

VCP: In the work of Lichtenwalter and Chawla (2012), proposed the concept of a vertex collocation profile (**VCP**) for the purpose of link analysis and prediction. In their definition, $VCP_{v,u}^{n,r}$ is a vector describing the relationship between two nodes v and u, in terms of their common membership in all possible subgraphs of n vertices over r relations. Figure 7 gives all possible elements that are included in $VCP_{s,t}^{3,2}$.

All of these methods are designed to describe the "proximity" between two nodes by capturing as much information as possible in heterogeneous networks. **Meta path** statistically measures the significance level of each meta path in supporting the likelihood of coauthorship for two authors. *MRLP* and *VCP* employ local substructures information to depict the link likelihood of two nodes in the future.

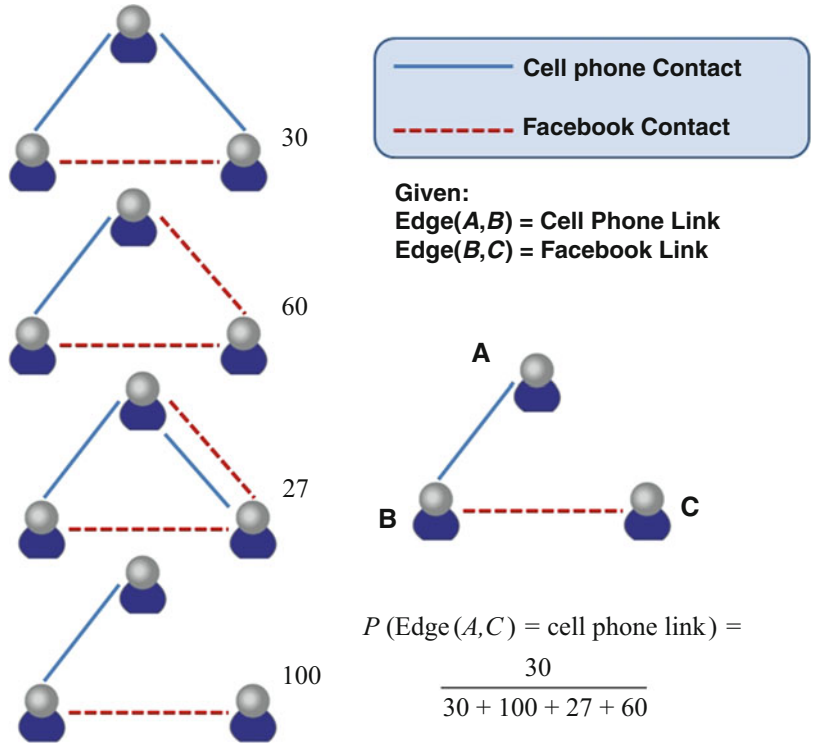
Link Prediction: A Primer, Table 1 Examples of meta path

Meta path	Description	Sig. level
$A - P \rightarrow P - A$	a_i cites a_j	**
$A - P \leftarrow P - A$	a_j cites a_i	***

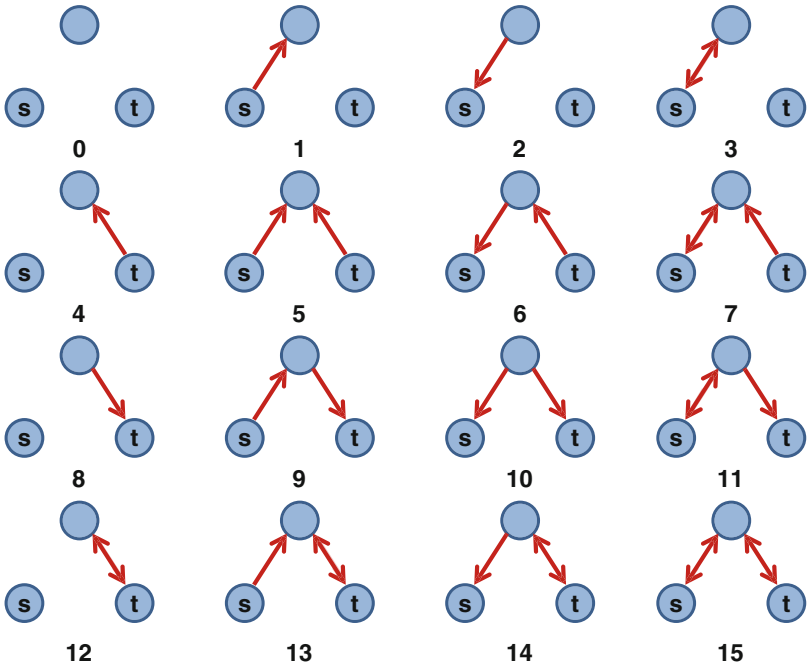
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$; ****$p < 0.001$

Key Applications

Beside friendship recommendation (Dong et al. 2012) in social networks, the link prediction has applications in many other domains. In the healthcare domain, the identification of



Link Prediction: A Primer, Fig. 6 MRLP toy example



Link Prediction: A Primer, Fig. 7 16 possible elements in $VCP_{s,t}^{3,2}$

interactions between drugs and target proteins or between diseases and proteins is a key area (Davis and Chawla 2011). In the area of e-commerce, one of the important usages of link prediction is to construct recommendation systems (Liu and Kou 2007).

Future Directions

Link prediction as a research area continues to evolve and is starting to play a key role in understanding and modeling network dynamics. Some of the recent work has included temporality in link prediction. Sun et al. (2012) developed the methodology for predicting when a link will form, not just whether it will form. Yang et al. (2012) showed that considering the temporal information significantly improves the performance of predicting new links. Recently cross-network prediction and transfer learning (Dong et al. 2012; Horvat et al. 2012) become hot topics in the link prediction field. Some of the open challenges remain in scaling the problem to include a much richer attribute space of links and nodes.

Cross-References

- ▸ Anonymization and De-anonymization of Social Network Data
- ▸ Distance and Similarity Measures
- ▸ Inferring Social Ties
- ▸ Social Recommendation in Dynamic Networks

References

Adamic LA, Adar E (2003) Friends and neighbors on the web. Soc Netw 25(3):211–230

Barabasi A-L, Jeong H, Neda Z, Ravasz E (2002) Evolution of the social network of scientific collaboration. Phys A 311(3–4):590–614

Clause A, Moore C, Newman MEJ (2008) Hierarchical structure and the prediction of missing links in network. Nature 453:98–101

Davis D, Chawla NV (2011) Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. PLoS One 6(7):e22670

Davis D, Lichtenwalter R, Chawla NV (2011) Multi-relational link prediction in heterogeneous information networks. In: ASONAM'11, Kaohsiung

Deng H, Han J, Zhao B, Yu Y, Lin CX (2011) Probabilistic topic models with biased propagation on heterogeneous information networks. In: KDD'11, San Diego

Dong Y, Tang J, Wu S, Tian J, Chawla NV, Rao J, Gao H (2012) Link prediction and recommendation across heterogeneous social networks. In: ICDM'12, Berlin

Horvat E-A, Hanselmann M, Hamprecht FA, Zweig KA (2012) One plus one makes three (for social networks). PLoS One 4:121–134

Kashima H, Abe N (2006) A parameterized probabilistic model of network evolution for supervised link prediction. In: ICDM'06, Leipzig

Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43

Kunegis J, Lommatzsch A (2009) Learning spectral graph transformations for link prediction. In: ICML'09, Montreal

Liben-Nowell D, Kleinberg J (2007) The link prediction problem for social networks. J Am Soc Inf Sci Technol 58(7):1019–1031

Lichtenwalter R, Chawla NV (2012) Vertex collocation profile: subgraph counting for link analysis and prediction. In: WWW'12, Lyon

Lichtenwalter R, Lussier J, Chawla NV (2010) New perspectives and methods in link prediction. In: KDD'10, Washington

Liu Y, Kou Z (2007) Predicting who rated what in large scale datasets. In: SIGKDD Exploration Newsletter 9:62–65

Menon AK, Elkan C (2011) Link prediction via matrix factorization. In: ECML-PKDD'11, Athens

Murata T, Moriyasu S (2007) Link prediction of social networks based on weighted proximity measures. In: IEEE/WIC/ACM ICWI'07, Vila Real

Newman MEJ (2001) Clustering and preferential attachment in growing networks. Phys Rev Lett E 64:025102

Qiu B, He Qi, Yen J (2011) Evolution of node behavior in link prediction. In: AAAI'11, San Francisco

Scellato S, Noulas A, Mascolo C (2011) Exploiting place features in link prediction on location-based social networks. In: KDD'11, San Diego

Sun Y, Barber R, Gupta M, Aggarwal CC, Han J (2011) Co-author relationship prediction in heterogeneous bibliographic networks. In: ASONAM'11, Kaohsiung

Sun Y, Han J, Aggarwal CC, Chawla NV (2012) When will it happen? Relationship prediction in heterogeneous information networks. In: WSDM'12, Seattle

Taskar B, Wong M-F, Abbeel P, Koller D (2003) Link prediction in relational data. In: NIP, Acapulco, Mexico

Viswanath B, Mislove A, Cha M, Gummadi KP (2009) On the evolution of user interaction in facebook. In: WSON'09, New York, NY, USA

Wang C, Satuluri V, Parthasarathy S (2007) Local probabilistic models for link prediction. In: ICDM'07, Omaha

Yang Y, Chawla NV, Sun Y, Han J (2012) Predicting links in multi-relational and heterogeneous networks. In: ICDM'12, Brussels

Link Rank

▸ Community Detection in Social Network: An Experience with Directed Graphs

Links

▸ Mapping Online Social Media Networks

Location

▸ Spatiotemporal Footprints in Social Networks

Location-Based Online Social Networks: Location-Based Online Social Media, Location-Based Online Social Services

▸ Privacy Preservation and Location-Based Online Social Networking

Location-Based Recommendation

▸ Spatiotemporal Personalized Recommendation of Social Media Content

Location-Based Social Networking Services

▸ Location-Based Social Networks