

Contents

Certificate	iii
Declaration by the Candidate	v
Copyright Transfer Certificate	vii
Preface	xi
Acknowledgements	xiii
Contents	xv
List of Figures	xxi
List of Tables	xxv
Abbreviations	xxvii
Symbols	xxix
1 Introduction	1
1.1 Complex networks	1
1.2 Link prediction	2
1.3 Evaluation criteria	4
1.4 Motivation of the thesis	4
1.5 Contribution of the thesis	5
1.5.1 A comprehensive experimental survey of link prediction	5
1.5.2 Influence of higher level clustering features towards link prediction	6
1.5.3 Influence of higher order paths indices towards link prediction . . .	7
1.6 Organization of the thesis	7

2	Social Networks Analysis Background	9
2.1	Social networks and graph Theory	9
2.2	General characteristics of social networks	10
2.2.1	Triadic closure and clustering coefficient	10
2.2.2	Small world phenomenon	11
2.2.3	Scale-free networks and preferential attachment	12
2.2.4	Homophily or assortative mixing	13
2.3	Common studied problems of social network analysis	14
2.3.1	Link-based object ranking (LBR)	14
2.3.2	Community detection	14
2.3.3	Finding central nodes	14
2.3.4	Influence maximization	15
2.3.5	Link prediction	15
2.4	Applications of social network analysis	16
	Application specific to link prediction	16
2.5	Computational complexity of social network analysis	17
	Computational Complexity of Similarity-based Link Prediction	18
3	Link Prediction Techniques, Applications, and Performance: A Survey	21
3.1	Similarity-based methods	22
3.1.1	Local similarity indices	23
3.1.1.1	Common neighbors (CN)	23
3.1.1.2	Jaccard coefficient (JC)	24
3.1.1.3	Adamic/Adar index (AA)	24
3.1.1.4	Preferential attachment (PA)	24
3.1.1.5	Resource allocation Index (RA)	25
3.1.1.6	Cosine similarity or Salton index (SI)	26
3.1.1.7	Sorensen index	26
3.1.1.8	CAR-based common neighbor index (CAR)	26
3.1.1.9	Hub promoted index (HPI)	27
3.1.1.10	Hub depressed index (HDI)	28
3.1.1.11	Local naive Bayes-based common neighbors (LNBCN)	28
3.1.1.12	Leicht-Holme-Newman local index (LHNL)	29
3.1.1.13	Node clustering coefficient (CCLP)	29
3.1.1.14	Node and link clustering coefficient (NLC)	30
3.1.2	Global similarity indices	30
3.1.2.1	Katz index	30
3.1.2.2	Random walk with restart (RWR)	31
3.1.2.3	Shortest path	32
3.1.2.4	Leicht-Holme-Newman global index (LHNG)	33
3.1.2.5	Cosine based on L^+ (Cos^+)	34

3.1.2.6	Average commute time (ACT)	34
3.1.2.7	Matrix forest index (MF)	35
3.1.2.8	SimRank (SR)	36
3.1.2.9	Rooted PageRank (RPR)	37
3.1.3	Quasi-local indices	37
3.1.3.1	Local path index (LP)	38
3.1.3.2	Path of length 3 (L3)	38
3.1.3.3	Similarity based on local random walk and superposed random walk (LRW and SRW)	40
	Remarks	41
3.2	Probabilistic and maximum likelihood models	41
3.2.1	Local probabilistic model for link prediction	42
3.2.2	Probabilistic relational model for link prediction (PRM)	44
3.2.3	Hierarchical structure model (HSM) [1]	45
3.2.4	Stochastic block model (SBM) [2]	47
3.2.5	Exponential random graph model (ERGM) or P-star model	50
3.3	Dimension reduction frameworks for link prediction	51
3.3.1	Embedding-based link prediction	52
3.3.2	Factorization-based frameworks for link prediction	55
3.4	Other approaches	57
3.4.1	Learning-based frameworks for link prediction	57
3.4.2	Information theory-based link prediction	58
3.4.3	Clustering-based link prediction	60
3.5	Experimental Setup and Results Analysis	62
3.5.1	Datasets	63
3.5.2	Accuracy	64
	Recall@k	65
	Area under the precision-recall curve (AUPR)	65
	Area under the receiver operating characteristics curve (AUROC)	66
	Average precision	68
	Concluding remarks	69
	Parameters settings	70
3.5.3	Efficiency	71
3.6	Variations of link prediction problem	73
3.6.1	Link prediction in weighted and directed networks	73
3.6.2	Link prediction in temporal networks	75
3.6.3	Link prediction in bipartite networks	76
3.6.4	Link prediction in heterogeneous networks	77
3.7	Link prediction applications	78
3.7.1	Network reconstruction	78

3.7.2	Recommender system	80
3.7.3	Network completion problem	81
3.7.4	Spam mail detection	82
3.7.5	Privacy control in social networks	82
3.7.6	Identifying missing references in a publication	83
3.7.7	Routing in networks	83
3.7.8	Incorporating user's influence in link prediction	84
4	Level-2 Node Clustering Coefficient-based Link Prediction	87
4.1	Introduction	87
4.2	Proposed work	89
	Extracting more local information	90
	Link prediction based on Level-2 node clustering coefficient.	92
	Algorithm description.	94
4.3	Experimental study	94
4.3.1	Evaluation metrics	94
4.3.2	Datasets description	96
4.3.3	Results analysis	98
	AUROC	98
	AUPR	99
	AP	102
	Recall	102
	Concluding remarks.	102
	Complexity analysis.	104
	Statistical test	106
4.4	Conclusion and future works	107
5	Link Prediction in Complex Networks Based on Significance of Higher-Order Path Index (SHOPI)	111
5.1	Introduction	111
5.2	Proposed work	115
	Significance of the path index of length 2	116
	Significance of path index of higher order	116
	Algorithm description	118
5.3	Experimental study	119
5.3.1	Evaluation metrics	119
5.3.2	Datasets description	122
5.3.3	Results analysis	123
	AUROC	123
	Average precision (AP)	125
	Effects of the parameter value ψ and sensitivity analysis	126
	AUROC sensitivity	126

AP sensitivity	126
Significance of higher order paths	129
Complexity analysis.	129
Statistical test	131
5.4 Conclusion	131
6 Conclusion and future directions	135
6.1 Conclusion	135
6.2 Future directions	136
Bibliography	139
A List of Publications	173

List of Figures

1.1	The Link Prediction (LP) finds missing links (i.e., AD, AC, and BD) in this observed network.	3
2.1	Triads in networks	10
2.2	Small world phenomenon with pure essential growth	12
3.1	Taxonomy of Link Prediction Approaches	22
3.2	$CAR\ Index = (Number\ of\ CNs) \times (Number\ of\ LCLs)$	27
3.3	Local probabilistic model for link prediction [3]	42
3.4	An illustrating example of HSM for a graph of 6 nodes and its two possible dendrograms as described in the paper [1]. The internal nodes of each dendrogram are labeled as the maximum likelihood probability \bar{p}_r , defined by the equation 3.45. The likelihoods of the left and the right dendrograms are $L(D_1) = (1/3)(2/3)^2 \cdot (1/4)^2(3/4)^6 = 0.00165$, and $L(D_2) = (1/9)(8/9)^8 = 0.0433$. Thus, the second (i.e., right) dendrogram is most probable as it divides the network in a balanced one at the first level.	48
3.5	The Karate club network (left) and its representation in the embedding space with the DeepWalk [4] algorithm.	52
3.6	Embedding of nodes x and y to the embedding space	53

3.7	An example illustrating the cycle formation link probability model [5], where the the probability of the missing link $(x - y)$ is generated by the following three mechanisms; random link occurrence $g(1)$, length-2 cycle generation $g(2)$ i.e. $(x - a - y, x - c - y)$, and length-4 cycle generation $g(3)$ i.e. $(x - b - d - y)$	61
4.1	Initially at time t_0 , three links are present in the disconnected graph. As the time progress, more links are formed as shown at the time instant $t_{(k-1)}$. Now, at the time instant t_k , which of the non-existing links (i.e. AC, AE, BC, BE, CD) will be formed ? Finding the potential links that will appear at t_k is called the link prediction problem.	88
4.2	Notion of level-2 clustering coefficient	90
4.3	Exploring local to global structure	91
4.4	Computing level-2 node clustering coefficient	92
4.5	AUROC Results	100
4.6	AUPR Results	101
4.7	AP Results	103
4.8	Recall Results	105
5.1	Path-based approaches to link prediction	113
5.2	Path length-2 score calculation: the score between x and y , $S(x, y) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$, and the score between p and q , $S(p, q) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$	116
5.3	Three possible paths (blue colored edges) between the node pair (x, y) during the computation of path length-3 score.	118
5.4	Confusion Matrix	120
5.5	AUROC results sensitivity corresponding to different parameter values of ψ	127

5.6 AP results sensitivity corresponding to different parameter values of ψ . . 128