Link Prediction in Complex Networks: A Survey

Ajay Kumar

Supervisor: Dr. Bhaskar Biswas

October 16, 2017

Outline

Introduction

- Overview
- State of the Art

Research Proposals

- Proposal 1.
- Proposal 2

Social Network

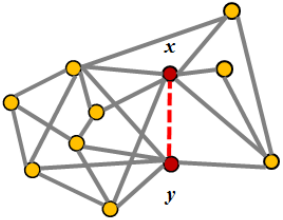
- Social network is a standard approach to model a communication in a group or community of persons.
- Such networks can be represented as graphical model in which
  - a node maps to a person or social entity, and
  - an edge corresponds to an association or collaboration between them.

Issues and Challenges

- The relationships among individuals are continuously changing so addition and/or deletion of several edges and vertices take place.
    - Results the social networks to be highly dynamic and complex.
- In case of pairwise classification problem, one of the fundamental challenges is dealing with the large outcome space; if there are n actors, there are n^2 possible choices to be taken care of.

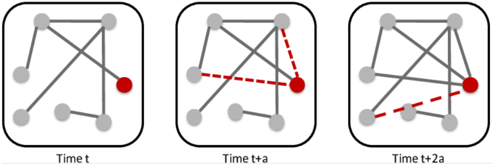
Link Prediction

- Here, we address a specific problem of social networks termed as Link Prediction (LP).
- Link Prediction can be defined in two scenarios:
  - In the first scenario **(also known as structural link prediction)**, given a snapshot of a network, infer which new interactions between nodes are likely to occur in the future [1].



Cont...

- In second scenario **(Temporal link prediction)**, given link data for times 1 through T, can we predict the links at time T + a, T+2a.... [2] ?



Formal Definition [1]

- Graphs/Network G = (V , E) where
  - V is the set of vertices in G, and E is the set of edges.
  - Consider a snapshot $G_{t_0-t_1}(V, E)$ of G during time interval $[t_0, t_1]$ and $E_{t_0-t_1}$ be the set of edges present in that snapshot.
  - The task of link prediction is to find set of edges $E_{t_0'-t_1'}$ during the time interval $[t_0', t_1']$ where $[t_0, t_1] \leq [t_0', t_1']$.

Applications

(a) Proposing items to users

(b) Friend recommendation

(c) Marriage proposals

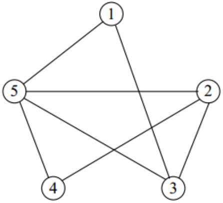
(d) spam emails detection

Evaluation Metrics

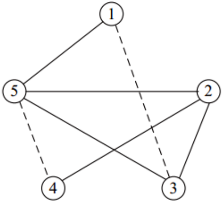
- To calculate the accuracy of algorithms, following metrics are used:
  - Area under the Receiver Operating Characteristics Curve (AUC) [3]
  - Precision [4,5]

Cont...

- consider a simple undirected network G(V; E) in which V is the set of nodes and E is the set of links.



Original network



Training network

Cont...

- The following observation can be made in the considered network:
  - Total possible links = U
  - Existent links = E
  - Non exitent links = $U - E$
  - Observed links = E^T = Training set
  - Non-Observed links = $U - E^T$
  - Missing links = E^P = Test set

Area under the ROC Curve

- **Area under the ROC Curve (AUC)**: Given a ranking of Non-observed links, the term AUC is estimated as the likelihood that a chosen missing link is given a higher score than a randomly chosen non-existent link.
- Each time two edges are selected randomly one from each set and compared their scores.
- Then, AUC can be calculated using the following expression:

$$AUC = \frac{n_1 + 0.5n_2}{n}$$

- where, n is total independent comparisons, n_1 is number of times the missing link with a higher score n_2 is number of times they have same score.

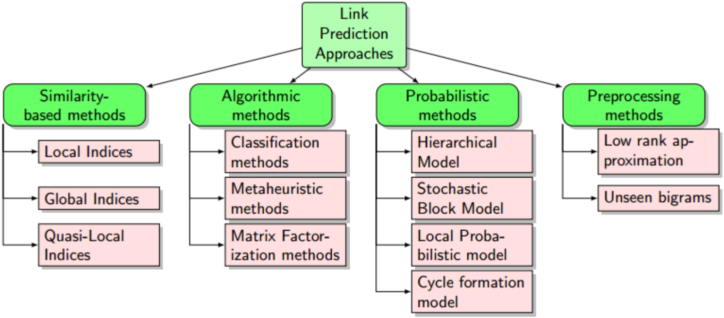
Precision

- **Precision**
  Given the ranking of non-observed links, precision can be defined as the proportion of relevant items to the number of items chosen. i.e.,

$$Precision = \frac{L_r}{L}$$

- where, L represents predicted links having top scores, and L_r, the number of predicted links which are correct.

Link Prediction Approaches: A Taxonomy



Link Prediction Approaches

Similarity-based methods
- Local Indices
- Global Indices
- Quasi-Local Indices

Algorithmic methods
- Classification methods
- Metaheuristic methods
- Matrix Factorization methods

Probabilistic methods
- Hierarchical Model
- Stochastic Block Model
- Local Probabilistic model
- Cycle formation model

Preprocessing methods
- Low rank approximation
- Unseen bigrams

State-of-the-Art

Framework	Model	Method Name	Expression	Reference
Similarity Based	Local	CN	$S_{xy} = \lvert\Gamma(x) \cap \Gamma(y)\rvert$	Newman M.E.J 2001
		Jaccard	$S_{xy} = \frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$	Jaccard P. 1901
		AA	$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log \lvert\Gamma(z)\rvert}$	Adamic and Adar 2003
		RA	$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\lvert\Gamma(z)\rvert}$	Zhou et al. 2009
		PA	$S_{xy} = K_x * K_y$	Barabasi and Albert 1999
		CAR	$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\lvert\Gamma(x) \cap \Gamma(y) \cap \Gamma(z)\rvert}{\lvert\Gamma(z)\rvert}$	Cannistraci et al. 2013
	Global	NSP	$S_{xy} = -\lvert shortestpath_{xy}\rvert$	Liben-Nowell 2005
		Katz Index	$S_{xy} = \sum_{l=1}^{\infty} \beta^l \lvert paths_{xy}^{<l>}\rvert = \sum_{l=1}^{\infty} \beta^l (A^l)_{xy}$	Leo Katz 1953
		Random Walk	$S_{xy} = \vec{P_y^x}(t) = P^T \vec{P_y^x}(t-1)$	Karl Pearson 1905
		Random Walk with Restart	$S_{xy} = q_{xy} + q_{yx}$ $q_x = cP^T \vec{q_x} + (1-c)\vec{e_x}$	Tong et al. 2006
		SimRank	$S_{xy} = \beta \frac{\sum_{i \in \Gamma_x} \sum_{j \in \Gamma_y} S(ij)}{\lvert\Gamma_x\rvert \lvert\Gamma_y\rvert}$	Jeh and Widom 2002
	Quasilocal	Local Path Index	$S_{xy} = A^2 + \epsilon A^3$	Lu et al. 2009
		LRW	$S_{xy}(t) = q_x \pi_{xy}(t) + q_y \pi_{yx}(t)$ $q_x = \frac{k_x}{M}$	Liu and Lu 2010
		SRW	$S_{xy}(t) = \sum_{\tau=1}^{t} [q_x \pi_{xy}(\tau) + q_y \pi_{yx}(\tau)]$	Liu and Lu 2010

Cont...

Framework	Method	Features and Characteristics	Model and Approach	Reference
Algorithmic Based	Classification	Topological, Aggregated and Proximity(shortest path)	DT, SVM, KNN, MLP	Al. Hasan et al. 2006
		Subgraph feature edge rank	Random Forest	Cuckierski et al. 2011
		Sum of patient, Ethinicity Sum of neighbors, Jaccard	SVM	Almansoori et al. 2012
	Metaheuristic	Heuristic function: CN, Fitness: deg-sum(path(i,j))	ACO	B. Chen et al. 2014
		Special subgraphs namely Bi-fan structure (4 nodes & 4 links)	ACO	E. Sherkat et al. 2014
		Linear combination of similarity indices and coefficient	Evolutionary strategy to optimize the coefficients	Bliss et al. 2014
	Factorization	shortest path (k=0,1,2)	Matrix factorization with bagging	Zhifeng Wu et al. 2016
		latent features with optional explicit features for nodes and edges	Matrix factorization	Menon and Elkan 2011
		communicability matrix $C_t = e^{\beta W_t}$	Symmetric NMF with Feature Collapsing algorithm	Xiaoke Ma et al. 2017

Proposal 1. Motivation

- **Motivation:** Wang and Go [6] proved that BA achieved results better than many other bio-inspired optimization techniques such as Ant Colony Optimization (ACO), Genetic Algorithm (GA), Harmony Search (HS), Particle Swarm Optimization (PSO), to solve numerical optimization problem.
- However, due to their stochastic nature of BA, swarm intelligence algorithms are never guaranteed to find an optimal solution for any problem, but they will often find a good solution if one exists.
- To mitigate this problem we can use chaos theory, in which generated sequences are well distributed.
- chaotic sequences perform well in escaping from local optimum.

BAT Framework: An Inroduction

- A nature inspired metaheuristic framework introduced by Xin-She Yang [7] in 2010.
- BAT framework works on echolocation behaviour of bats.
- Microbats use echolocation to detect prey, avoid obstacles, and locate their roosting crevices in the dark.
- These bats emit a very loud sound pulse and listen for the echo that bounces back from the surrounding objects.
- With the help of variance in these pulse properties, bats decide their hunting strategy.
- Frequency of bats ranges from 25kHz to 150kHz

Problem formulation

- Ling Chen et al.[8] performed link prediction based on direct optimization of area under the ROC curve (AUCD) as

  $$AUC = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I(w^T x_i > w^T x_j)$$

  where, m and n are total number of positive class and negative class examples. I(.) is an indicator function defined as

  $$I(q) = \begin{cases} 1 & \text{if q is true} \\ 0 & \text{otherwise} \end{cases}$$

- We have used the above concept in our bat framework as objective funcion.

BAT algorithm steps

- Step 1: Set objective function
  - Our objective is to minimize L(w) using BAT algorithm

$$L(w) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} max[0, 1 - w^T(x_i - x_j)] + \frac{\lambda}{2} ||w||_2^2$$

- Step 2: Parameter initialization
  - Position (or solution) X: binary string of length N ($features \in [0, 1]$).
  - Velocity V: change in features.
  - Frequency (f) and Loudness (A) are initialized by Gause map.
  - Pulse rate emission (r) is initialized by tent map.
  - Set number of iterations (t) and current best solution (x_*).

Cont...

- Step 3: Evaluate fitness of all bats and compare with x_*.

$$x_{new} = x_{old} + \epsilon.A^t$$

- Step 4: Update the V, f, x, A and r.

$$f_i = f_{min} + (f_{max} - f_{min})\beta$$
$$v_i^t = v_i^{t-1} + (x_* - x_i^t)f_i$$
$$x_i^t = x_i^{t-1} + v_i^t$$
$$A_i^{t+1} = \alpha.A_i^t$$
$$r_i^{t+1} = r_i^0[1 - \exp(-\gamma.t)]$$

  Here, $\alpha(\alpha \in [0,1])$ and $\gamma(\gamma > 0)$ are positive constants, $\beta(\beta \in [0,1])$ is a random number in a uniform distribution.

- Step 5: Iterate the procedure until the maximum number of iteration or convergence.

Proposal 2. Motivation

- **Motivation:** Most link prediction algorithms are based on topological properties of a network varies from local to global.
- Very less work have been done on **structural identity** of nodes in the network [9].
- **structural identity** is a concept of symmetry in which nodes are identified by network structure.
- LFR Ribeiro [9] considers structural identity to represent nodes of a network and prove it to be scalable for large networks, which might be very useful in link prediction problem.

Link prediction through learning node representation from structural identity

- **Structural similarity:** Two nodes that have same degree are structurally similar, but if their neighbors also have the same degree, then they are even more structurally similar.
- **Step 1.** Determine **structural similarity** between each vertex pair in the graph for different neighborhood sizes.
- The structural distance between vertices a and b having k-hop distant (i.e. k-neighbohood) is

$$f_k(a,b) = f_{k-1}(a,b) + g[ODS(N_k(a)), ODS(N_k(b))]$$

$$k \geq 0, |N_k(a)|, |N_k(b)| \geq 0, f_{-1} = 0$$

Cont...

- Here, $|N_k(a)|$ and $|N_k(b)|$ are sets of nodes at k-hop distant respectively. $g(D_1, D_2) \geq 0$ measures the distance between ordered degree sequences(ODSs) D_1 and D_2.
- Then Dynamic Time Warpping (DTW) [10] is used to compare two ordered degree sequences of same or different sizes.
- Given a local distance measure d, DTW computes the optimal alignment between two sequences having minimal cost.
- The local distance function depends on the dimension of the feature representation.

Cont...

- For example, in case of 1-dimensional feature

$$d(a, b) = \frac{max(a, b)}{min(a, b)} - 1$$

- For a 2-dimensional feature, the Manhattan distance can be applied for this purpose.

$$d(a, b) = |a - b|$$

- **Step 2.** Now, our objective is to optimize the alignment using some optimization algorithm.

References

1. Liben-Nowell, D., Kleinberg, J. The link prediction problem for social networks. Journal of the American Society for Information Science and Technology 58(7), 2007, pp. 1019-1031.
2. Daniel M. Dunlavy, Tamara G. Kolda, Evrim Acar. ”Temporal Link Prediction using Matrix and Tensor Factorizations”, ACM Transactions on Knowledge Discovery from Data 5(2):10 (27 pages), February 2011.
3. J. A. Hanely, B.J.M: The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology 143(1), 1982, pp. 29-36.
4. Geisser, S: Predictive inference: An introduction. Chapman and Hall, New York (1993).
5. J. L. Herlocker J. A. Konstann, K.T.J.T.R.: Evaluating collaborative filtering recommender systems,. ACM Trans. Inf. Syst. (22 (2004) 5).
6. Gaige Wang and Lihong Guo, A Novel Hybrid Bat Algorithm with Harmony Search for Global Numerical Optimization. Journal of Applied Mathematics, 2013.
7. X.-S. Yang, A New Metaheuristic Bat-Inspired Algorithm, in: Nature Inspired Cooperative Strategies for Optimization (NICSO 2010) (Eds. J. R. Gonzalez et al.), SCI 284, 65-74 (2010).
8. Caiyan Dai, Ling Chen, Bin Li. Network link prediction based on direct optimization of area under curve. Applied Intelligence, pp 427437, 2017.
9. L.F.R. Ribeiro, P.H.P Saverese, D.R. Figueiredo. struc2vec: Learning Node Representations from Structural Identity, KDD2017.
10. S Salvador and P Chan. FastDTW: Toward accurate dynamic time warping in linear time and space. In Workshop on Min. Temp. and Seq. Data, ACM SIGKDD 2004.