Supervised Learning for Link Prediction Using Similarity Indices

Sergey Korolev[1] and Leonid Zhukov

Higher School of Economics, Moscow, Russia,
sokorolev@edu.hse.ru

Abstract. The abstract should summarize the contents of the paper using at least 70 and at most 150 words. It will be set in 9-point font size and be inset 1.0 cm from the right and left margins. There will be two blank lines before and after the Abstract. ...

Keywords: network analysis, graph theory, link prediction

1 Introduction

Over the last few years the topic of link prediction has become very popular due to the rise of recommendation systems and social networks. The

The aim of this paper is to try to devise an approach to predict the missing links in the network using only structural information of said network.

2 Similarity Indices

In this section, we will describe most popular local and global similarity indices and measure their accuracy in ranking links according to the probability that said links were removed from the network.

We'll start with describing the method for measuring said accuracy. We start with the set of edges E of existing graph G. Then we shuffle it and split it into n parts $E'_i,\ i \in \{1, \dots, n\}$. Then for each step we subtract E'_i from E and look at the complement of the remaining set of edges $\bar{E}_i$ to the set of edges of complete graph on these nodes, so that $E^* = E \cup \bar{E} = E_i \cup E'_i \cup \bar{E} = E_i \cup \bar{E}_i$, where E^* is the set of edges of complete graph and $\bar{E}$ is complement of graph G.

The main tool for measuring the accuracy of ranking is AUC score, which can be described as the probability that for pair of edges (e_a, e_b) $e_a \in E'_i,\ e_b \in \bar{E}$, the ranking is such that $score(e_a) > score(e_b)$ plus 0.5 times the probability that $score(e_a) = score(e_b)$. Or as formula – $AUC = \frac{n' + 0.5n''}{n}$, where n is number of pairs (e_a, e_b) $e_a \in E'_i,\ e_b \in \bar{E}$, n' is number of pairs such that $score(e_a) > score(e_b)$ and n'' is number of pairs such that $score(e_a) = score(e_b)$.

We'll first describe all indices used for the prediction of links and then compare their AUC scores on different networks.

2.1 Local Indices

Common neighbours This is the easiest and most obvious approach to the idea of capturing the similarity of two nodes in the network. As such, a number of indices described below will use this measure with different weights. So if we denote $\Gamma(x)$ the number of neighbours of node x in the graph, the formula looks like

$$s_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|,$$

where s_{xy} is the index.

Salton Index This index is also called cosine similarity and is defined as

$$s_{xy}^{Salton} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x \times k_x}},$$

where k_x is the degree of node x.

Jaccard Index

$$s_{xy}^{Jaccard} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}.$$

Sørensen Index

$$s_{xy}^{Sørensen} = \frac{2|\Gamma(x) \cap \Gamma(y)|}{k_x + k_y}.$$

Hub Promoted Index (HPI)

$$s_{xy}^{HPI} = \frac{|\Gamma(x) \cap \Gamma(y)|}{min\{k_x, k_y\}}.$$

Hub Depressed Index (HDI)

$$s_{xy}^{HDI} = \frac{|\Gamma(x) \cap \Gamma(y)|}{max\{k_x, k_y\}}.$$

LeichtHolmeNewman Index (LHN1)

$$s_{xy}^{LHN1} = \frac{|\Gamma(x) \cap \Gamma(y)|}{k_x \times k_y}.$$

Preferential Attachment Index (PA)

$$s_{xy}^{PA} = k_x \times k_y.$$

AdamicAdar Index (AA)

$$s_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{logk_z}.$$

Resource Allocation Index (RA)

$$s_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}.$$

2.2 Global Similarity Indices

Katz Index

$$S^{Katz} = (I - \beta A)^{-1} - I.$$

LeichtHolmeNewman Index (LHN2)

$$S^{LHN2} = 2m\lambda_1 D^{-1}(I - \frac{\phi A}{\lambda_1})^{-1} D^{-1}.$$

Average Commute Time (ACT)

$$S_{xy}^{ACT} = \frac{1}{l_{xx}^{+} + l_{yy}^{+} - 2l_{xy}^{+}}.$$

Cosine based on L^{+}

$$s_{xy}^{cos^{+}} = cos(x, y)^{+} = \frac{v_x^T v_y}{|v_x| \cdot |v_y|} = \frac{l_{xy}^{+}}{\sqrt{l_{xx}^{+} \cdot l_{yy}^{+}}}.$$

Random Walk with Restart (RWR)

$$s_{xy}^{RWR} = q_{xy} + q_{yx}.$$

Matrix Forest Index (MFI)

$$S = (I + L)^{-1}.$$

Table 1. AUCs of local indices. AN - AdjNoun network, CN - CelegansNeural network, Dp - Dolphins network, FB - Football network, LM - LesMiserables network, PB - PolBooks network, KC - Karate club network, UA - USAir network

	CN	SaI	JI	SoI	HPI	HDI	LHN1	PAI	AAI	RAI
AN	0.662	0.606	0.603	0.603	0.618	0.603	0.566	0.744	0.662	0.659
CN	0.844	0.797	0.790	0.790	0.805	0.779	0.725	0.750	0.861	0.866
Dp	0.779	0.774	0.779	0.779	0.763	0.780	0.762	0.619	0.781	0.781
FB	0.846	0.856	0.858	0.858	0.856	0.857	0.859	0.270	0.846	0.846
LM	0.910	0.882	0.880	0.880	0.847	0.878	0.820	0.776	0.918	0.919
PB	0.887	0.884	0.875	0.875	0.894	0.863	0.848	0.653	0.897	0.890
KC	0.700	0.636	0.607	0.607	0.712	0.593	0.600	0.712	0.726	0.733
UA	0.934	0.908	0.897	0.897	0.869	0.891	0.767	0.885	0.945	0.951

Table 2. AUCs of global indices.

	KI	LHN2	ACT	CBL	RWR	MFI
AN	0.714	0.549	0.742	0.586	0.738	0.667
CN	0.852	0.717	0.738	0.848	0.725	0.865
Dp	0.799	0.825	0.760	0.791	0.649	0.804
FB	0.857	0.879	0.588	0.885	0.273	0.878
LM	0.884	0.813	0.863	0.825	0.794	0.867
PB	0.891	0.858	0.729	0.891	0.618	0.899
KC	0.755	0.610	0.666	0.739	0.618	0.749
UA	0.920	NaN	0.892	0.913	0.862	0.913

3 Method

We begin by defining the test dataset to check the performance of the algorithm. We shuffle the set of existing edges and split it into n parts $E'_i,\ i \in \{1, \ldots, n\}$. It is said that $n = 10$ achieves the best complexity-precision tradeoff. Now we can use the set $\bar{E}_i = E'_i \cup \bar{E}$ as test dataset, where if $e \in E'_i$ then $class_e = 1$ and if $e \in \bar{E}$ then $class_e = 0$.

Now we need to construct training dataset to train our classifier on it and then check its performance on the test dataset. We take the remaining set of edges E_i, shuffle it and split it into n parts $(E'_i)'_j$. To achieve balanced training dataset with the same amount of edges with class 0 and 1 we split the complement $\bar{E}_i$ into subsets of the same size as $(E'_i)'_j$. Now for each subset $(E'_i)'_j$ we take this subset out of E_i, then calculate indices on the complement of the remaining set and mark classes as $class_e$, where if $e \in (E'_i)'_j$ then $class_e = 1$ and if $e \in \bar{E}_i$ then $class_e = 0$. And to get balanced dataset, on each step we take only edges from one subset of $\bar{E}_i$ as examples of class 0.

After n steps of the above process we get the trainig dataset of size $2|E_i|$. Now we can train our classifier on this dataset and then test its performance on the test dataset. After testing SVM, decision trees, k nearest neighbours, random forests, LDA and QDA it was observed that the best performance is achieved with classification using random forests ().

4 Results

Table 3. Scores of the algorithm performance on the class of marked links.

	F1	Precision	Recall	ROCAUC	Accuracy
AN	0.027 ± 0.003	0.014 ± 0.002	0.579 ± 0.072	0.638 ± 0.032	0.696 ± 0.023
CN	0.041 ± 0.001	0.021 ± 0.000	0.811 ± 0.018	0.809 ± 0.008	0.808 ± 0.004
Dp	0.049 ± 0.008	0.025 ± 0.005	0.635 ± 0.082	0.704 ± 0.040	0.771 ± 0.035
FB	0.199 ± 0.025	0.116 ± 0.017	0.721 ± 0.044	0.831 ± 0.019	0.939 ± 0.011
LM	0.163 ± 0.021	0.090 ± 0.013	0.831 ± 0.068	0.875 ± 0.032	0.918 ± 0.013
PB	0.080 ± 0.009	0.042 ± 0.005	0.780 ± 0.080	0.812 ± 0.040	0.842 ± 0.012
KC	0.092 ± 0.026	0.049 ± 0.014	0.641 ± 0.145	0.717 ± 0.082	0.790 ± 0.043
UA	0.083 ± 0.008	0.044 ± 0.004	0.901 ± 0.024	0.910 ± 0.011	0.919 ± 0.009

We assume that H is (A_∞, B_∞)-subquadratic at infinity, for some constant symmetric matrices A_∞ and B_∞, with $B_\infty - A_\infty$ positive definite. Set:

$$\gamma := \text{smallest eigenvalue of } \ B_\infty - A_\infty \tag{1}$$

$$\lambda := \text{largest negative eigenvalue of } \ J\frac{d}{dt} + A_\infty \ . \tag{2}$$

Theorem 1 tells us that if $\lambda + \gamma < 0$, the boundary-value problem:

$$\begin{aligned} \dot{x} &= JH'(x) \\ x(0) &= x(T) \end{aligned} \tag{3}$$

has at least one solution $\overline{x}$, which is found by minimizing the dual action functional:

$$\psi(u) = \int_o^T \left[\frac{1}{2}\left(\Lambda_o^{-1}u, u\right) + N^*(-u)\right] dt \tag{4}$$

on the range of Λ, which is a subspace $R(\Lambda)_L^2$ with finite codimension. Here

$$N(x) := H(x) - \frac{1}{2}\left(A_\infty x, x\right) \tag{5}$$

is a convex function, and

$$N(x) \leq \frac{1}{2}\left(\left(B_\infty - A_\infty\right)x, x\right) + c \quad \forall x \ . \tag{6}$$

Proposition 1. *Assume $H'(0) = 0$ and $H(0) = 0$. Set:*

$$\delta := \liminf_{x \to 0} 2N(x) \left\|x\right\|^{-2} \ . \tag{7}$$

If $\gamma < -\lambda < \delta$, the solution $\overline{u}$ is non-zero:

$$\overline{x}(t) \neq 0 \quad \forall t \ . \tag{8}$$

Proof. Condition (7) means that, for every $\delta' > \delta$, there is some $\varepsilon > 0$ such that

$$\left\|x\right\| \leq \varepsilon \Rightarrow N(x) \leq \frac{\delta'}{2}\left\|x\right\|^2 \ . \tag{9}$$

It is an exercise in convex analysis, into which we shall not go, to show that this implies that there is an $\eta > 0$ such that

$$f\left\|x\right\| \leq \eta \Rightarrow N^*(y) \leq \frac{1}{2\delta'}\left\|y\right\|^2 \ . \tag{10}$$

Since u_1 is a smooth function, we will have $\left\|hu_1\right\|_\infty \leq \eta$ for h small enough, and inequality (10) will hold, yielding thereby:

$$\psi(hu_1) \leq \frac{h^2}{2}\frac{1}{\lambda}\left\|u_1\right\|_2^2 + \frac{h^2}{2}\frac{1}{\delta'}\left\|u_1\right\|^2 \ . \tag{11}$$

Fig. 1. This is the caption of the figure displaying a white eagle and a white horse on a snow field

If we choose δ' close enough to δ, the quantity $\left(\frac{1}{\lambda} + \frac{1}{\delta'}\right)$ will be negative, and we end up with

$$\psi(hu_1) < 0 \qquad \text{for} \quad h \neq 0 \quad \text{small} \; . \tag{12}$$

On the other hand, we check directly that $\psi(0) = 0$. This shows that 0 cannot be a minimizer of ψ, not even a local one. So $\overline{u} \neq 0$ and $\overline{u} \neq \Lambda_o^{-1}(0) = 0$. □

Corollary 1. *Assume H is C^2 and (a_∞, b_∞)-subquadratic at infinity. Let ξ_1, $\ldots, \xi_N$ be the equilibria, that is, the solutions of $H'(\xi) = 0$. Denote by ω_k the smallest eigenvalue of $H''(\xi_k)$, and set:*

$$\omega := \text{Min}\, \{\omega_1, \ldots, \omega_k\} \; . \tag{13}$$

If:

$$\frac{T}{2\pi} b_\infty < -E\left[-\frac{T}{2\pi} a_\infty\right] < \frac{T}{2\pi}\omega \tag{14}$$

then minimization of ψ yields a non-constant T-periodic solution $\overline{x}$.

We recall once more that by the integer part $E[\alpha]$ of $\alpha \in \mathbb{R}$, we mean the $a \in \mathbb{Z}$ such that $a < \alpha \leq a + 1$. For instance, if we take $a_\infty = 0$, Corollary 2 tells us that $\overline{x}$ exists and is non-constant provided that:

$$\frac{T}{2\pi} b_\infty < 1 < \frac{T}{2\pi} \tag{15}$$

or

$$T \in \left(\frac{2\pi}{\omega}, \frac{2\pi}{b_\infty}\right) \; . \tag{16}$$

Proof. The spectrum of Λ is $\frac{2\pi}{T}\mathbb{Z} + a_\infty$. The largest negative eigenvalue λ is given by $\frac{2\pi}{T} k_o + a_\infty$, where

$$\frac{2\pi}{T} k_o + a_\infty < 0 \leq \frac{2\pi}{T}(k_o + 1) + a_\infty \; . \tag{17}$$

Hence:

$$k_o = E\left[-\frac{T}{2\pi} a_\infty\right] \; . \tag{18}$$

The condition $\gamma < -\lambda < \delta$ now becomes:

$$b_\infty - a_\infty < -\frac{2\pi}{T}k_o - a_\infty < \omega - a_\infty \tag{19}$$

which is precisely condition (14). □

Lemma 1. *Assume that H is C^2 on $\mathbb{R}^{2n}\backslash\{0\}$ and that $H''(x)$ is non-degenerate for any $x \neq 0$. Then any local minimizer $\widetilde{x}$ of ψ has minimal period T.*

Proof. We know that $\widetilde{x}$, or $\widetilde{x} + \xi$ for some constant $\xi \in \mathbb{R}^{2n}$, is a T-periodic solution of the Hamiltonian system:

$$\dot{x} = JH'(x) \ . \tag{20}$$

There is no loss of generality in taking $\xi = 0$. So $\psi(x) \geq \psi(\widetilde{x})$ for all $\widetilde{x}$ in some neighbourhood of x in $W^{1,2}\left(\mathbb{R}/T\mathbb{Z}; \mathbb{R}^{2n}\right)$.

But this index is precisely the index $i_T(\widetilde{x})$ of the T-periodic solution $\widetilde{x}$ over the interval $(0, T)$, as defined in Sect. 2.6. So

$$i_T(\widetilde{x}) = 0 \ . \tag{21}$$

Now if $\widetilde{x}$ has a lower period, T/k say, we would have, by Corollary 31:

$$i_T(\widetilde{x}) = i_{kT/k}(\widetilde{x}) \geq k i_{T/k}(\widetilde{x}) + k - 1 \geq k - 1 \geq 1 \ . \tag{22}$$

This would contradict (21), and thus cannot happen. □

Notes and Comments. The results in this section are a refined version of [1]; the minimality result of Proposition 14 was the first of its kind.

To understand the nontriviality conditions, such as the one in formula (16), one may think of a one-parameter family x_T, $T \in \left(2\pi\omega^{-1}, 2\pi b_\infty^{-1}\right)$ of periodic solutions, $x_T(0) = x_T(T)$, with x_T going away to infinity when $T \to 2\pi\omega^{-1}$, which is the period of the linearized system at 0.

Table 4. This is the example table taken out of *The TEXbook,* p. 246

Year	World population
8000 B.C.	5,000,000
50 A.D.	200,000,000
1650 A.D.	500,000,000
1945 A.D.	2,300,000,000
1980 A.D.	4,400,000,000

Theorem 1 (Ghoussoub-Preiss). *Assume $H(t,x)$ is $(0,\varepsilon)$-subquadratic at infinity for all $\varepsilon > 0$, and T-periodic in t*

$$H(t,\cdot) \qquad \text{is convex} \quad \forall t \tag{23}$$

$$H(\cdot,x) \qquad \text{is} \quad T-\text{periodic} \quad \forall x \tag{24}$$

$$H(t,x) \geq n\left(\|x\|\right) \qquad \text{with} \quad n(s)s^{-1} \to \infty \quad \text{as} \quad s \to \infty \tag{25}$$

$$\forall \varepsilon > 0 \ , \quad \exists c \ : \ H(t,x) \leq \frac{\varepsilon}{2}\|x\|^2 + c \ . \tag{26}$$

Assume also that H is C^2, and $H''(t,x)$ is positive definite everywhere. Then there is a sequence x_k, $k \in \mathbb{N}$, of kT-periodic solutions of the system

$$\dot{x} = JH'(t,x) \tag{27}$$

such that, for every $k \in \mathbb{N}$, there is some $p_o \in \mathbb{N}$ with:

$$p \geq p_o \Rightarrow x_{pk} \neq x_k \ . \tag{28}$$

□

*Example 1 (*External forcing*).* Consider the system:

$$\dot{x} = JH'(x) + f(t) \tag{29}$$

where the Hamiltonian H is $(0, b_\infty)$-subquadratic, and the forcing term is a distribution on the circle:

$$f = \frac{d}{dt}F + f_o \qquad \text{with} \quad F \in L^2\left(\mathbb{R}/T\mathbb{Z};\mathbb{R}^{2n}\right) \ , \tag{30}$$

where $f_o := T^{-1}\int_o^T f(t)dt$. For instance,

$$f(t) = \sum_{k\in\mathbb{N}} \delta_k \xi \ , \tag{31}$$

where δ_k is the Dirac mass at $t = k$ and $\xi \in \mathbb{R}^{2n}$ is a constant, fits the prescription. This means that the system $\dot{x} = JH'(x)$ is being excited by a series of identical shocks at interval T.

Definition 1. *Let $A_\infty(t)$ and $B_\infty(t)$ be symmetric operators in $\mathbb{R}^{2n}$, depending continuously on $t \in [0,T]$, such that $A_\infty(t) \leq B_\infty(t)$ for all t.*

A Borelian function $H : [0,T] \times \mathbb{R}^{2n} \to \mathbb{R}$ is called (A_∞, B_∞)-subquadratic at infinity if there exists a function $N(t,x)$ such that:

$$H(t,x) = \frac{1}{2}\left(A_\infty(t)x, x\right) + N(t,x) \tag{32}$$

$$\forall t \ , \quad N(t,x) \qquad \text{is convex with respect to} \quad x \tag{33}$$

$$N(t,x) \geq n\left(\|x\|\right) \qquad \text{with} \quad n(s)s^{-1} \to +\infty \quad \text{as} \quad s \to +\infty \tag{34}$$

$$\exists c \in \mathbb{R} \ : \quad H(t,x) \leq \frac{1}{2}\left(B_\infty(t)x, x\right) + c \quad \forall x \ . \tag{35}$$

If $A_\infty(t) = a_\infty I$ and $B_\infty(t) = b_\infty I$, with $a_\infty \leq b_\infty \in \mathbb{R}$, we shall say that H is (a_∞, b_∞)-subquadratic at infinity. As an example, the function $\|x\|^\alpha$, with $1 \leq \alpha < 2$, is $(0,\varepsilon)$-subquadratic at infinity for every $\varepsilon > 0$. Similarly, the Hamiltonian

$$H(t,x) = \frac{1}{2}k\|k\|^2 + \|x\|^\alpha \tag{36}$$

is $(k, k+\varepsilon)$-subquadratic for every $\varepsilon > 0$. Note that, if $k < 0$, it is not convex.

Notes and Comments. The first results on subharmonics were obtained by Rabinowitz in [5], who showed the existence of infinitely many subharmonics both in the subquadratic and superquadratic case, with suitable growth conditions on H'. Again the duality approach enabled Clarke and Ekeland in [2] to treat the same problem in the convex-subquadratic case, with growth conditions on H only.

Recently, Michalek and Tarantello (see [3] and [4]) have obtained lower bound on the number of subharmonics of period kT, based on symmetry considerations and on pinching estimates, as in Sect. 5.2 of this article.

References

1. Clarke, F., Ekeland, I.: Nonlinear oscillations and boundary-value problems for Hamiltonian systems. Arch. Rat. Mech. Anal. 78, 315–333 (1982)
2. Clarke, F., Ekeland, I.: Solutions périodiques, du période donnée, des équations hamiltoniennes. Note CRAS Paris 287, 1013–1015 (1978)
3. Michalek, R., Tarantello, G.: Subharmonic solutions with prescribed minimal period for nonautonomous Hamiltonian systems. J. Diff. Eq. 72, 28–55 (1988)
4. Tarantello, G.: Subharmonic solutions for Hamiltonian systems via a $\mathbb{Z}_p$ pseudoindex theory. Annali di Matematica Pura (to appear)
5. Rabinowitz, P.: On subharmonic solutions of a Hamiltonian system. Comm. Pure Appl. Math. 33, 609–633 (1980)