Journal of Computational Science 5 (2014) 750–764

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

An evolutionary algorithm approach to link prediction in dynamic social networks

Catherine A. Bliss*, Morgan R. Frank, Christopher M. Danforth, Peter Sheridan Dodds

Computational Story Lab, Department of Mathematics and Statistics, Vermont Complex Systems Center & the Vermont Advanced Computing Core, University of Vermont, Burlington, VT 05405, United States

ARTICLE INFO

Article history:
Received 26 April 2013
Received in revised form 4 January 2014
Accepted 12 January 2014
Available online 5 February 2014

Keywords:
Algorithms
Data mining
Link prediction
Social networks
Twitter
Complex networks
Complex systems

ABSTRACT

Many real world, complex phenomena have underlying structures of evolving networks where nodes and links are added and removed over time. A central scientific challenge is the description and explanation of network dynamics, with a key test being the prediction of short and long term changes. For the problem of short-term link prediction, existing methods attempt to determine neighborhood metrics that correlate with the appearance of a link in the next observation period. Recent work has suggested that the incorporation of topological features and node attributes can improve link prediction. We provide an approach to predicting future links by applying the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) to optimize weights which are used in a linear combination of sixteen neighborhood and node similarity indices. We examine a large dynamic social network with over 10^6 nodes (Twitter reciprocal reply networks), both as a test of our general method and as a problem of scientific interest in itself. Our method exhibits fast convergence and high levels of precision for the top twenty predicted links. Based on our findings, we suggest possible factors which may be driving the evolution of Twitter reciprocal reply networks.

1. Introduction

Time varying social networks can be used to model groups whose dynamics change over time. Individuals, represented by nodes, may enter or exit the network, while interactions, represented by links, may strengthen or weaken. Most network growth models capture global properties, but do not capture specific localized dynamics such as who will be connected to whom in the future. And yet, it is precisely this type of information that would be most valuable in applications such as national security, online social networking sites (people you may know), and organizational studies (predicting potential collaborators).

In this paper, we focus primarily on the link prediction problem: given a snapshot of a network $G_t = (V, E_t)$, with nodes V (nodes present across all time steps) and links E_t, at time t, we seek to predict the most likely links to newly occur in the next timestep, $t+1$ [1].

Link prediction strategies may be broadly categorized into three groups: similarity based strategies, maximum likelihood algorithms, and probabilistic models. As noted by Lu et al. [2], the latter two approaches can be prohibitively time consuming for a large network over 10,000 nodes. Given our interest in large, sparse networks with $N \gtrsim 10^6$, we focus primarily on local information and use similarity indices to characterize the likelihood of future interactions. We consider the two major classes of similarity indices: topological-based and node attribute (Table 1).

There does not appear to be one best similarity index that is superior in all settings. Depending on the network under analysis, various measures have shown to be particularly promising [1,3–8]. These findings suggest that the predictors which work "best" for a given network may be related to the inherent structure within the individual network rather than a universal best set of predictors. Further, it is also plausible that the best link predictor may change as the network responds to endogenous and exogenous factors driving its evolution.

Topological similarity indices encode information about the relative overlap between nodes' neighborhoods. We expect that the more "similar" two nodes' topological neighborhoods are (e.g., the more overlap in their shared friends), the more likely they may be to exhibit a future link. The common neighbors index, a building block of many other topological similarity indices, has been shown

* Corresponding author.
E-mail addresses: Catherine.Bliss@uvm.edu (C.A. Bliss), Morgan.Frank@uvm.edu (M.R. Frank), Chris.Danforth@uvm.edu (C.M. Danforth), pdodds@uvm.edu (P.S. Dodds).

1877-7503/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jocs.2014.01.003

Table 1
The sixteen similarity indices chosen for inclusion in the link predictor. We define the *neighborhood of node u* to be $\Gamma(u) = \{v \in V | e_{u,v} \in E\}$, where $G = (V, E)$ is a network, consisting of vertices (V) and edges (E). The degree of node u is represented by k_u, the adjacency matrix is denoted by A, and a path of length n between $u, v \in V$ is denoted as $\mathcal{P}_n(u, v)$.

Topological similarity indices (abbreviation)		
Jaccard Index (J)	$J(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)\cup\Gamma(v)\rvert}$	Measures the probability that a neighbor of u or v is a neighbor of both u and v. This measurement is a way of characterizing shared content and has been shown to be meaningful in information retrieval [15]
Adamic–Adar coefficient (A)	$A(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{\log(\lvert\Gamma(z)\rvert)}$	Quantifies features shared by nodes u and v and weights rarer features more heavily [19]. Interpreting this in the context of neighborhoods, the Adamic–Adar coefficient can be used to characterize neighborhood overlap between nodes u and v, weighting the overlap of smaller such neighborhoods more heavily
Common neighbors (C)	$C(u,v) = \lvert\Gamma(u)\cap\Gamma(v)\rvert$	Measures the number of shared neighbors between u and v. Despite the simplicity of this index, Newman [9] documented that the probability of future links occurring in a collaboration network was positively correlated with the number of common neighbors
Average path weight (P)	$P(u,v) = \frac{\sum_{p\in\mathcal{P}_2(u,v)\cup\mathcal{P}_3(u,v)} w_p}{\lvert\mathcal{P}_2(u,v)\rvert+\lvert\mathcal{P}_3(u,v)\rvert}$	Computes the sum of the minimum weights on the directed paths between u and v divided by the number of paths between u and v, where only paths of lengths 2 and 3 are considered due to the large size of this network. We take w_p to be the minimum weight of the edges in the path, in the spirit that a path's strength is only as strong as its weakest edge
Katz (K)	$K = \sum_{n=1}^{\infty} \beta^n A^n$	Computed as such, the Katz is a global index [20]. This series converges to $(I-\beta A)^{-1} - I$, when $\beta < \max(\lambda(A))$. When $\beta \ll 1$ then K approximates the number of common neighbors. Due to the size of our network and computational expense of this index, we truncate to $n = 3$. We set $\beta = 1$ because we are not concerned with convergence &to emphasize the number of paths of length greater than two. Previous observations suggest that individuals who appear to be connected by a path length of n in Twitter RRNs may actually be connected by a path of shorter length due to role of missing data [34]
Preferential attachment (Pr)	$Pr(u,v) = k_u \times k_v$	Gives higher scores to pairs of nodes for which one or both have high degree. This index arose from the observation that nodes in some networks acquire new links with a probability proportional to their degree [9] and preferential attachment random growth models [10]
Resource allocation (R)	$R(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{\lvert\Gamma(z)\rvert}$	Considers the amount of a given resource one node has and assumes that each node will distribute its resource equally among all neighbors [3]
Hub promoted index (Hp)	$Hp(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\min\{k_u,k_v\}}$	First proposed to measure the topological overlap of pairs of substrates in metabolic networks, this index assigns higher scores to links adjacent to hubs since the denominator depends on the minimum degree of the two users [11]
Hub depressed index (Hd)	$Hd(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\max\{k_u,k_v\}}$	When one of the nodes has large degree, the denominator will be larger and thus *Hd* is smaller in the case where one of the users is a hub [13]
Leicht–Holme–Newman index (L)	$L(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{k_u k_v}$	Measures the number of common neighbors relative to the square of their geometric mean. This index gives high similarities to pairs of nodes that have many common neighbors compared to the expected number of such neighbors [14]
Salton index (Sa)	$Sa(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\sqrt{k_u k_v}}$	Measures the number of common neighbors relative to their geometric mean [15]
Sorenson index (So)	$So(u,v) = \frac{2\lvert\Gamma(u)\cap\Gamma(v)\rvert}{k_u+k_v}$	Measures the number of common neighbors relative to their arithmetic mean. This index is similar to *J*, however *J* counts the number of (unique) nodes in the shared neighborhood. This index was previously used to establish equal amplitude groups in plant sociology based on the similarity of species [16]
Individual characteristics similarity indices		
Id similarity (I)	$I(u,v) = 1 - \frac{\lvert Id(u)-Id(v)\rvert}{\max\{\lvert Id(a)-Id(b)\rvert\}_{a,b\in V}}$	In 2008, user ids were numbered sequentially and a user's id served as a proxy for the relative length of time since opening a Twitter account. Id similarity characterizes the extent to which two individuals adopt Twitter simultaneously
Tweet count similarity (T)	$T(u,v) = 1 - \frac{\lvert T(u)-T(v)\rvert}{\max\{\lvert T(a)-T(b)\rvert\}_{a,b\in V}}$	Tweet count $T(u)$ measures the number of tweets we have gathered for node u in a given week. Tweet count similarity quantifies how similar two individuals' tweet counts are, with 1 representing identical tweet counts and 0 representing dissimilar tweet counts
Happiness similarity (H)	$H(u,v) = 1 - \frac{\lvert h(u)-h(v)\rvert}{\max\{\lvert h(a)-h(b)\rvert\}_{a,b\in V}}$	Building on previous work [40], happiness scores ($h(u)$ and $h(v)$) are computed as the average of happiness scores for words authored by users u and v during the week of analysis
Word similarity (W)	$W(u,v) = 1 - \frac{1}{2}\sum_{n=1}^{50,000} \lvert f_{u,n} - f_{v,n}\rvert$	From a corpus consisting of the 50,000 most commonly occurring words used in Twitter from 2008 through 2011 [40], the similarity of words used by u and v is computed by a modified Hamming distance, where $f_{u,n}$ represents the normalized frequency of word usage of the nth word by user u. The value of $W(u,v)$ ranges from 0 (dissimilar word usage) to 1 (similar word usage) [34]

to correlate with the occurrence of future links [9]. Several variants of this index have been proposed and have been shown to be useful for link prediction in a variety of settings [3,10–18]. See [2] for a review. In their seminal paper on link prediction, Liben-Nowell and Kleinberg [1] examined author collaboration networks derived from arXiv submissions in four subfields of Physics. They found that neighborhood similarity measures, such as the Jaccard [15], Adamic–Adar [19], and the Katz coefficients [20] provided a large factor improvement over randomly predicted links.

As a complement for topological similarity indices, node-specific similarity indices examine node attributes, such as language, topical similarity, and behavior, in the case of social networks. Several studies have suggested that incorporating these measures can enhance link prediction [2,4,22–26]. In training algorithms for link prediction, researchers have used supervised learning including support vector machine [27], decision trees [4], bagged random forests [17], supervised random walks [6], multilayer perceptrons, and others. Notably, Al Hasan et al. [27] use both