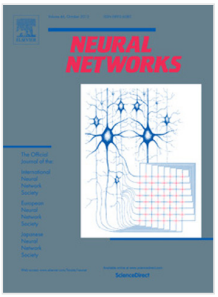
Accepted Manuscript



A bag-of-paths framework for network data analysis

Kevin Françoisse, Ilkka Kivimäki, Amin Mantrach, Fabrice Rossi, Marco Saerens

PII: S0893-6080(17)30066-7
DOI: http://dx.doi.org/10.1016/j.neunet.2017.03.010
Reference: NN 3736

To appear in: *Neural Networks*

Received date: 5 July 2016
Revised date: 17 March 2017
Accepted date: 23 March 2017

Please cite this article as: Françoisse, K., Kivimäki, I., Mantrach, A., Rossi, F., & Saerens, M. A bag-of-paths framework for network data analysis. *Neural Networks* (2017), http://dx.doi.org/10.1016/j.neunet.2017.03.010

A Bag-of-Paths Framework for Network Data Analysis

Kevin Françoisse[a], Ilkka Kivimäki[a,c], Amin Mantrach[d], Fabrice Rossi[e], Marco Saerens[a,b]

[a] *Université catholique de Louvain, Belgium*
[b] *Université Libre de Bruxelles, Belgium*
[c] *Aalto University, Department of Computer Science, Helsinki, Finland*
[d] *Yahoo! Research, Sunnyvale, California, USA*
[e] *Université Paris 1 Panthéon-Sorbonne, France*

Abstract

This work develops a generic framework, called the bag-of-paths (BoP), for link and network data analysis. The central idea is to assign a probability distribution on the set of all paths in a network. More precisely, a Gibbs-Boltzmann distribution is defined over a bag of paths in a network, that is, on a representation that considers all paths independently. We show that, under this distribution, the probability of drawing a path connecting two nodes can easily be computed in closed form by simple matrix inversion. This probability captures a notion of relatedness, or more precisely accessibility, between nodes of the graph: two nodes are considered as highly related when they are connected by many, preferably low-cost, paths. As an application, two families of distances between nodes are derived from the BoP probabilities. Interestingly, the second distance family interpolates between the shortest-path distance and the commute-cost distance. In addition, it extends the Bellman-Ford formula for computing the shortest-path distance in order to integrate sub-optimal paths (exploration) by simply replacing the minimum operator by the soft minimum operator. Experimental results on semi-supervised classification tasks show that both of the new distance families are competitive with other state-of-the-art approaches. In addition to the distance measures studied in this paper, the bag-of-paths framework enables straightforward computation of many other relevant network measures.

Keywords: Network science, link analysis, distance and similarity on a graph, shortest-path distance, resistance distance, commute-time distance, semi-supervised classification.

1. Introduction

1.1. General introduction

Network and link analysis is a highly studied field, subject of much recent work in various areas of science: applied mathematics, computer science, social science, physics, chemistry, pattern recognition, applied statistics, data mining and machine learning, to name a few [4, 21, 32, 60, 66, 78, 93, 103, 109]. Within this context, one key issue is the proper quantification of the structural relatedness between nodes of a network by taking both direct and indirect connections into account [68]. This problem is faced in all disciplines involving networks in