Physica A 463 (2016) 293–302

Contents lists available at ScienceDirect

Physica A

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

# Traveling salesman problems with PageRank Distance on complex networks reveal community structure

Zhongzhou Jiang, Jing Liu*, Shuai Wang

*Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an 710071, China*

## HIGHLIGHTS

- Community detection problems are transformed to traveling salesman problems (TSPs).
- A distance measure based on PageRank is designed in TSPs.
- A threshold-based method is designed to transform the tours in TSPs to communities.
- A TSP-based community detection method, termed as TSP-CDA, is proposed.
- Networks with up to 10,000 nodes and varying structures are used in experiments.

## ARTICLE INFO

*Article history:*
Received 29 March 2016
Received in revised form 29 June 2016
Available online 26 July 2016

*Keywords:*
Community detection problems
Traveling salesman problems
PageRank

## ABSTRACT

In this paper, we propose a new algorithm for community detection problems (CDPs) based on traveling salesman problems (TSPs), labeled as TSP-CDA. Since TSPs need to find a tour with minimum cost, cities close to each other are usually clustered in the tour. This inspired us to model CDPs as TSPs by taking each vertex as a city. Then, in the final tour, the vertices in the same community tend to cluster together, and the community structure can be obtained by cutting the tour into a couple of paths. There are two challenges. The first is to define a suitable distance between each pair of vertices which can reflect the probability that they belong to the same community. The second is to design a suitable strategy to cut the final tour into paths which can form communities. In TSP-CDA, we deal with these two challenges by defining a PageRank Distance and an automatic threshold-based cutting strategy. The PageRank Distance is designed with the intrinsic properties of CDPs in mind, and can be calculated efficiently. In the experiments, benchmark networks with 1000–10,000 nodes and varying structures are used to test the performance of TSP-CDA. A comparison is also made between TSP-CDA and two well-established community detection algorithms. The results show that TSP-CDA can find accurate community structure efficiently and outperforms the two existing algorithms.

## 1. Introduction

Complex networks have been used in many fields to represent various kinds of complex systems [1,2]. To understand and utilize the information in complex networks, researchers have found many distinctive network properties like the small-world and scale-free ones [3–5], and developed various methods to capture structures and characteristics of networks from different perspectives, where the research on analyzing the community structure has drawn a great deal of attention during

* Corresponding author.
*E-mail addresses:* neouma@mail.xidian.edu.cn, neouma@163.com (J. Liu).

http://dx.doi.org/10.1016/j.physa.2016.07.050
0378-4371/© 2016 Elsevier B.V. All rights reserved.

the past decade [6–27]. The purpose of community detection problems (CDPs) is to identify the modules and, possibly, their hierarchical organization, by only using the information encoded in the graph topology.

Although a great number of community detection algorithms have been proposed, the history of the research on this area is still not so long compared to the research on another NP-hard problem, Traveling Salesman Problems (TSPs). The TSP was proposed in the 1800s and first formulated in 1930. TSPs are one of the most intensively studied problems in the field of optimization, and have been used as a benchmark for various types of optimization methods. The purpose of TSPs is to find a tour with minimum cost over a set of cities, which visits each city exactly once, and comes back to the starting city. The distance between each pair of cities is predefined and can be used as the cost function. Although TSPs are computationally difficult, a large number of heuristics and exact methods have been designed, and some of them can tackle instances with tens of thousands of cities [28–33].

Since TSPs need to find a tour with minimum cost, cities close to each other are usually clustered in the tour. This inspired us to model CDPs as TSPs by thinking each vertex as a city. Then, in the final tour, the vertices in the same community are clustered together, and the community structure can be obtained by cutting the tour into a couple of paths. Therefore, we propose a TSP-based CD algorithm, termed as TSP-CDA. The first challenge is to define a suitable distance between each pair of vertices, which can be used as the cost function in TSPs. To let vertices in the same community be clustered in the final tour, this distance should manifest network structures; that is, the distance should indicate the likelihood of two vertices belonging to the same community. Thus, borrowing the idea of PageRank algorithm [34], we first designed a PageRank Feature (PRF) for each vertex and then based on the PRF, a PageRank Distance (PRD) is designed for each pair of vertices. The second challenge lies in cutting the tour into paths so that each path corresponds to a community. We designed a threshold based strategy to convert the tour into separated communities.

In the experiments, the performance of TSP-CDA is systematically tested on benchmark synthetic networks with 1000–10 000 nodes. A comparison is also made between our method and two well-established algorithms. The experimental results show that TSP-CDA outperforms the other algorithms.

The rest of this paper is organized as follows: Section 2 gives a brief introduction on the related work. Section 3 introduces the PageRank feature we designed. The details of TSP-CDA are given in Section 4. The experiments are reported in Section 5. Finally, Section 6 summarizes the work in this paper.

## 2. Related work

During the past decade, the research on analyzing the community structure in complex networks has drawn a great deal of attention, and various kinds of algorithms have been proposed. Girvan et al. in Ref. [6] proposed the Girvan–Newman (GN) algorithm which is one of the most known algorithms proposed so far. In Ref. [8], Newman proposed the well-know measure “modularity $Q$” to evaluate the quality of obtained communities. There are also other studies on community identification from complex networks that utilize physics-based methods.

Rosvall et al. [15] introduced an information theoretic approach to reveal community structures in weighted and directed networks. The probability flow of random walks on a network is used as a proxy for information flows in the real system and decomposes the network into modules by compressing a description of the probability flow. The result is a map that both simplifies and highlights the regularities in the structure and their relationship.

Lancichinetti et al. [16] presented the Order Statistics Local Optimization Method (OSLOM), which is the first method capable to detect clusters in network accounting for edge directions, edge weights, overlapping communities, hierarchies, and community dynamics. It is based on the local optimization of a fitness function expressing the statistical significance of clusters with respect to random fluctuations, which is estimated with tools of Extreme and Order Statistics. OSLOM can be used alone or as a refinement procedure of partitions/covers delivered by other techniques.

Meo et al. [17] proposed a strategy to enhance existing CD algorithms by adding a pre-processing step in which edges are weighted according to their centrality, w.r.t. the network topology. In this approach, the centrality of an edge reflects its contribution to make arbitrary graph transversals. i.e., spreading messages over the network, as short as possible. The proposed strategy is able to effectively complement information about network topology and can be used as an additional tool to enhance community detection. The computation of edge centralities is carried out by performing multiple random walks of bounded length on the network.

## 3. PageRank feature

The PageRank algorithm [34] is a well known website ranking algorithm used by the Google search engine. The PageRank value actually indicates the probability of a web page visited by a user. Suppose there is a website network with $N$ pages $p_1, p_2 \dots p_N$, and a user is visiting websites by jumping from one page to another using their hyperlinks. Then, for a specific page $p_i$, the probability that $p_i$ is visited after $t$ jumps, labeled as $PR^t(p_i)$, is

$$PR^t(p_i) = \sum_{j=1}^{N} L(p_j, p_i) PR^{t-1}(p_j) \tag{1}$$