Fast Katz and Commuters: Efficient Estimation of Social Relatedness in Large Networks

Pooya Esfandiar[1], Francesco Bonchi[2], David F. Gleich[3], Chen Greif[1], Laks V.S. Lakshmanan[1], and Byung-Won On[1]

[1] University of British Columbia, Vancouver BC, Canada
{pooyae,greif,laks,bwon}@cs.ubc.ca
[2] Yahoo! Research, Barcelona, Spain
bonchi@yahoo-inc.com
[3] Sandia National Laboratories*, Livermore CA, USA
dfgleic@sandia.gov

Abstract. Motivated by social network data mining problems such as link prediction and collaborative filtering, significant research effort has been devoted to computing topological measures including the Katz score and the commute time. Existing approaches typically approximate all pairwise relationships simultaneously. In this paper, we are interested in computing: the score for a single pair of nodes, and the top-k nodes with the best scores from a given source node. For the pairwise problem, we apply an iterative algorithm that computes upper and lower bounds for the measures we seek. This algorithm exploits a relationship between the Lanczos process and a quadrature rule. For the top-k problem, we propose an algorithm that only accesses a small portion of the graph and is related to techniques used in personalized PageRank computing. To test the scalability and accuracy of our algorithms we experiment with three real-world networks and find that these algorithms run in milliseconds to seconds without any preprocessing.

1 Introduction

The availability of large social networks and social interaction data (on movies, books, music, etc) have caused people to ask: what can we learn by mining this wealth of data? Measures of social relatedness play a fundamental role in answering this question. For example, Liben-Nowell and Kleinberg [13] identify a variety of topological measures as features for *link prediction*, the problem of predicting the likelihood of users/entities forming social ties in the future, given the current state of the network. The measures they studied fall into two categories – neighborhood-based measures and path-based measures. The former are cheaper to compute, yet the latter are more effective at link prediction. Katz

* Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

R. Kumar and D. Sivakumar (Eds.): WAW 2010, LNCS 6516, pp. 132–145, 2010.
© Springer-Verlag Berlin Heidelberg 2010

scores [11] were among the best link predictors, and the commute time [6] also performed well. Other uses of Katz scores and commute time are anomalous link detection [18], recommendation [20], and clustering [19].

Katz scores measure the affinity between nodes via a weighted sum of the number of paths between them. Formally, the Katz score between node i and j is $K_{i,j} = \sum_{\ell=1}^{\infty} \alpha^{\ell} \text{paths}_{\ell}(x, y)$, where $\text{paths}_{\ell}(x, y)$ denotes the number of paths of length ℓ between i to j and $\alpha < 1$ is an attenuation parameter. Let A be the symmetric adjacency matrix, and recall that $(A^{\ell})_{i,j}$ is the number of paths between node i and j. Then for all pairs of nodes,

$$K = \alpha A + \alpha^2 A^2 + \cdots = (I - \alpha A)^{-1} - I,$$

where the series converges if $\alpha < 1/\|A\|_2$.

The hitting time from node i to j is the expected number of steps for a random walk started at i to visit j, and the commute time between nodes is defined as the sum of hitting times from i to j and from j to i. The hitting time may be expressed using the row-stochastic transition matrix P with first-transition analysis: $H_{i,i} = 0$ and $H_{i,j} = 1 + \sum_k P_{i,k} H_{\ell,j}$. Unlike Katz, hitting time is not symmetric; but commute time is by definition, since $C = H + H^T$. Computing H and C via these definitions is not straightforward, and using the graph Laplacian, $L = D - A$ where D is the diagonal matrix of degrees, provides another means of computing the commute time. With the Laplacian, $C_{i,j} = \text{Vol}(G)(L^{\dagger}_{i,i} - 2L^{\dagger}_{i,j} + L^{\dagger}_{j,j})$ where $\text{Vol}(G)$ is the sum of elements in A and $L^{\dagger}$ is the pseudo-inverse of L [5].

Computing both of these measures between all pairs of nodes involves inverting a matrix, i.e. $(I - \alpha A)^{-1}$ or $L^{\dagger}$. Standard algorithms for a matrix inverse require $O(n^3)$ time and $O(n^2)$ memory and are inappropriate for a large network (see Section 2 for a brief survey of existing alternatives). Inspired by applications in anomalous link detection and recommendation [18,20], we focus on computing only a single Katz score or commute time and on computing the k most related nodes by Katz score.

In Section 3, we propose customized methods for the pairwise problems based on the Lanczos/Stieltjes procedure [8]. We specialize it for the Katz and commute time measures, providing a novel and useful application for the Lanczos/Stieltjes procedure. In Section 4, we present an algorithm to approximate the strongest ties between a given source node and its neighbors in terms of the Katz score (while we discuss the case of commute time in the conclusion section). This algorithm are inspired by a technique for personalized PageRank computing [14,2,3], though heavily adapted to the Katz score. We evaluate these methods on three real-world networks and report the results in Section 5. Our methods produce answers in seconds or milliseconds, whereas preprocessing techniques may often take over 10 minutes.

We have made our codes and data available for others to reproduce our results: `http://stanford.edu/~dgleich/publications/2010/codes/fast-katz/`.