Direction-Aware Proximity on Graphs

Hanghang Tong
Carnegie Mellon University, USA

Yehuda Koren
AT&T Labs - Research, USA

Christos Faloutsos
Carnegie Mellon University, USA

INTRODUCTION

In many graph mining settings, measuring node proximity is a fundamental problem. While most of existing measurements are (implicitly or explicitly) designed for undirected graphs; edge directions in the graph provide a new perspective to proximity measurement: measuring the proximity from A to B; rather than between A and B. (See Figure 1 as an example).

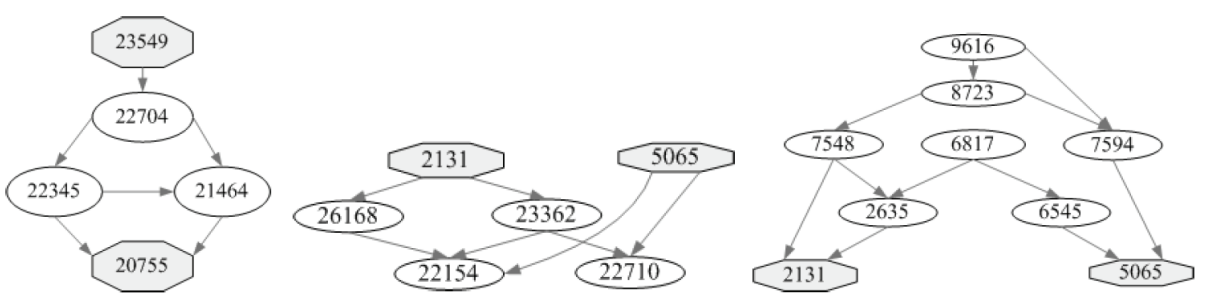
In this chapter, we study the role of edge direction in measuring proximity on graphs. To be specific, we will address the following fundamental research questions in the context of direction-aware proximity:

1. **Problem definitions:** How to define a direction-aware proximity?
2. **Computational issues:** How to compute the proximity score efficiently?
3. **Applications:** How can direction-aware proximity benefit graph mining?

BACKGROUND

In the literature, there are several measures of node proximity. Most standard measures are based on basic graph theoretical concepts -the shortest path length and the maximum flow. However the dependency of these measures on a single element of the graph, the shortest path or the minimum cut, makes them more suitable to managed networks, but inappropriate for measuring the random nature of relationships within social networks or other self organizing networks. Consequently, some works suggested more involved measures such as the sink-augmented delivered current (Faloutsos, Mccurley & Tomkins, 2004), cycle free effective conductance

Figure 1. An example of Dir-CePS. Each node represents a paper and edge denotes 'cited-by' relationship. By employing directional information, Dir-CePS can explore several relationships among the same query nodes (the two octagonal nodes): (a) A query-node to query-node relations; (b) common descendants of the query nodes (paper number 22154 apparently merged the two areas of papers 2131 and 5036); (c) common ancestors of the query nodes: paper 9616 seems to be the paper that initiated the research areas of the query papers/nodes. See 'Main Focus' section for the description of Dir-CePS algorithm.



(a) $\vec{f} = [+1, -1]$ (b) $\vec{f} = [+1, +1]$ (c) $\vec{f} = [-1, -1]$

(Koren, North & Volinsky, 2006), survivable network (Grötschel, Monma & Stoer, 1993), random walks with restart (He, Li, zhang, Tong & Zhang, 2004; Pan, Yang, Faloutsos & Duygulu, 2004). Notice that none of the existing methods meets all the three desirable properties that our approach meets: (a) dealing with directionality, (b) quality of the proximity score and (c) scalability.

Graph proximity is an important building block in many graph mining settings. Representative work includes connection subgraph (Faloutsos, McCurley & Tomkins, 2004; Koren, North & Volinsky, 2006; Tong & Faloutsos 2006), personalized PageRank (Haveliwala, 2002), neighborhood formulation in bipartite graphs (Sun, Qu, Chakrabarti & Faloutsos, 2005), content-based image retrieval (He, Li, zhang, Tong & Zhang, 2004), cross modal correlation discovery (Pan, Yang, Faloutsos & Duygulu, 2004), the BANKS system (Aditya, Bhalotia, Chakrabarti, Hulgeri, Nakhe & Parag 2002), link prediction (Liben-Nowell & Kleinberg, 2003), detecting anomalous nodes and links in the graph (Sun, Qu, Chakrabarti & Faloutsos, 2005), ObjectRank (Balmin, Hristidis & Papakonstantinou, 2004) and RelationalRank (Geerts, Mannila & Terzi, 2004).

MAIN FOCUS

In this Section, we begin by proposing a novel direction-aware proximity definition, based on the notion of escape probability of random walks. It is carefully designed to deal with practical problems such as the inherent noise and uncertainties associated with real life networks. Then, we address computational efficiency, by concentrating on two scenarios: (1) the computation of a single proximity on a large, disk resident graph (with possibly millions of nodes). (2) The computation of multiple pairwise proximities on a medium sized graph (with up to a few tens of thousand nodes). For the former scenario, we develop an iterative solution to avoid matrix inversion, with convergence guarantee. For the latter scenario, we develop an efficient solution, which requires only a single matrix inversion, making careful use of the so-called block-matrix inversion lemma. Finally, we apply our direction-aware proximity to some real life problems. We demonstrate some encouraging results of the proposed direction-aware proximity for predicting the existence of links together with their direction. Another application is so-called "directed center-piece subgraphs."

Direction-Aware Proximity: Definitions

Here we give the main definitions behind our proposed node-to-node proximity measure, namely, the escape probability; then we give the justification for our modifications to it.

Escape Probability

Following some recent works (Faloutsos, McCurley & Tomkins, 2004; Koren, North & Volinsky, 2006; Tong & Faloutsos 2006), our definition is based on properties of random walks associated with the graph. Random walks mesh naturally with the random nature of the self-organizing networks that we deal with here. Importantly, they allow us to characterize relationships based on multiple paths. Random walk notions are known to parallel properties of corresponding electric networks (Doyle & Snell, 1984). For example, this was the basis for the work of Faloutsos et al. (Faloutsos, McCurley & Tomkins, 2004) that measured nodes proximity by employing the notion of effective conductance. Since electric networks are inherently undirected, they cannot be used for our desired directed proximity measure. Nonetheless, the effective conductance can be adequately generalized to handle directional information by using the concept of escape probability (Doyle & Snell, 1984):

DEFINITION 1. The escape probability from node i to node j, $ep_{i,j}$, is the probability that the random particle that starts from node i will visit node j before it returns to node i.

Thus we adopt the escape probability as the starting point for our direction-aware node-to-node proximity score. That is, for the moment, we define the proximity $\text{Prox}(i, j)$ from node i to node j as exactly $ep_{i,j}$.

An important quantity for the computation of $ep_{i,j}$ is the *generalized voltage* at each of the nodes, denoted by $v_k(i, j)$: this is defined as the probability that a random particle that starts from node k will visit node j before node i. This way, our proximity measure can be stated as: