

Better Approximation of Betweenness Centrality*

Robert Geisberger[†]

Peter Sanders[†]

Dominik Schultes[†]

Abstract

Estimating the importance or centrality of the nodes in large networks has recently attracted increased interest. *Betweenness* is one of the most important centrality indices, which basically counts the number of shortest paths going through a node. Betweenness has been used in diverse applications, e.g., social network analysis or route planning. Since exact computation is prohibitive for large networks, approximation algorithms are important. In this paper, we propose a framework for unbiased approximation of betweenness that generalizes a previous approach by Brandes. Our best new schemes yield significantly better approximation than before for many real world inputs. In particular, we also get good approximations for the betweenness of unimportant nodes.

1 Introduction

One of the most important aspects of automatic analysis of networks is the computation of *centrality indices* that measure the importance of a node in some well defined way. Recently, the focus of attention in network analysis has shifted to the analysis of ever larger networks that are rapidly becoming available in such diverse areas as transportation networks (e.g., public transportation or road networks), social networks (e.g., friendship circles, recommendation networks, or citation networks), computer networks (e.g., the internet or peer-to-peer networks), or networks in bioinformatics (e.g., protein interaction networks).

In this paper we consider *betweenness centrality* [8, 1], which is one of the most frequently considered centrality indices. Our results might also be applicable to related concepts such as *stress centrality* [15] that are also based on counting shortest paths. Consider a weighted directed (multi)-graph $G = (V, E)$ with $n = |V|$, $m = |E|$. Let SP_{st} denote the set of shortest paths between source s and target t and $SP_{st}(v)$ the subset of SP_{st} consisting of paths that have v in their

interior. Then, the betweenness centrality for node v is

$$c_B(v) := \sum_{s,t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \quad (1.1)$$

where $\sigma_{st} := |SP_{st}|$ and $\sigma_{st}(v) := |SP_{st}(v)|$.

This definition counts the number of shortest paths through v , counting paths with alternatives only fractionally.

Our original motivation for considering betweenness was to identify sets of important nodes that can define a highway-node hierarchy [14], which is used for (dynamic) routing in road networks. For this application, the requirement is to process huge networks with many million nodes in a few minutes. In this context, we also need reasonable approximations for the betweenness of *all* nodes since we have to decide which of several neighboring unimportant nodes will make it to the first level of the hierarchy.

1.1 Related Work. Brandes [4] gives an exact algorithm for computing betweenness of all nodes that is based on solving a single source shortest path problem (SSSP) from each node. An SSSP computation from s produces a directed acyclic graph (DAG) encoding all shortest paths starting at s . By backward aggregation of counter values, the contributions of these paths to the betweenness counters can be computed in linear time (Section 3 gives more details). Depending on the graph model, the exact algorithm takes time between $\Theta(nm)$ (e.g., for unit edge weights) and $\Theta(nm + n^2 \log(n))$ (comparison based general edge weights). Although this is polynomial time, it is prohibitive for networks with many millions of nodes and edges. Bader and Madduri [2] present a massively parallel implementation of the exact algorithm that can handle a few million nodes.

Brandes and Pich [5] investigate how the exact algorithm can be turned into an approximation algorithm by extrapolating from a subset of k starting nodes (*pivots*), otherwise using the same aggregation strategy as the exact algorithm. Subsequently, we refer to this approximation algorithm as “Brandes’ algorithm”. A random sample of starting nodes turns out to work well. In particular, the randomized method yields an unbi-

*Partially supported by DFG grant SA 933/1-3.

[†]Universität Karlsruhe (TH), 76128 Karlsruhe, Germany, {robert.geisberger, sanders, schultes}@ira.uka.de

