k-Centralities: Local Approximations of Global Measures Based on Shortest Paths

Jürgen Pfeffer Kathleen M. Carley

Institute of Software Research, School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
+1 (412) 268-2418

jpfeffer@cs.cmu.edu kathleen.carley@cs.cmu.edu

ABSTRACT

A lot of centrality measures have been developed to analyze different aspects of importance. Some of the most popular centrality measures (e.g. betweenness centrality, closeness centrality) are based on the calculation of shortest paths. This characteristic limits the applicability of these measures for larger networks. In this article we elaborate on the idea of bounded-distance shortest paths calculations. We claim criteria for *k*-centrality measures and we introduce one algorithm for calculating both betweenness and closeness based centralities. We also present normalizations for these measures. We show that *k*-centrality measures are good approximations for the corresponding centrality measures by achieving a tremendous gain of calculation time and also having linear calculation complexity Θ(n) for networks with constant average degree. This allows researchers to approximate centrality measures based on shortest paths for networks with millions of nodes or with high frequency in dynamically changing networks.

Categories and Subject Descriptors

G.2.2 [**Discrete Mathematics**]: Graph Theory – *Graph algorithms*.

General Terms

Algorithms, Measurement, Performance.

Keywords

Large networks; Centrality approximation; Shortest paths; Betweenness centrality; Closeness centrality.

1. INTRODUCTION

The idea of *centrality* has a long tradition in the analysis of networks [6] [30]. Centrality calculations try to identify the most *important* nodes and are part of almost any network analytical project nowadays. Different centrality measures [43] cover different interpretations of *importance*. Centrality measures fit very well to the idea of ordering social actors upon their importance. Another reason for the popularity of centrality measures is their reduction of multi-dimensional complex network data into one dimensional information with just one value per node. This makes it possible to analyze network data with standard statistical algorithms and tools. Questions of the correlation of centrality and attributes of actors (e.g. age, gender, income) can be discussed.

The centrality of a node is a function of its position within a given network. Therefore the direct and indirect connections of a node to all other nodes play an essential role in the calculation of centrality measures. Some measures are calculated by looking at all direct connections of a node, e.g. degree centrality [27], or by also including all possible indirect paths into the calculation, e.g. information centrality [42]. Other Measures are based on shortest paths, e.g. betweenness centrality [1] [26] or closeness centrality [27]. A common limitation for all algorithms based on shortest paths is the calculation time needed for larger networks. For example, without the use of parallelization techniques [2], it is almost impossible to calculate the most between actors of a large scale online social networks like Facebook and Twitter. But researchers already work with this data, e.g. [45]. Other sources for huge network data are patents [5], wikis [31], or communication networks [35]. However, dealing with large networks is not just a problem for a small group of researchers working on supercomputers. The improvements in both computer speed and storage capacity enable almost every researcher or student to collect and store a great amount of data on their personal computer at universities or at home. Therefore, a big need for centrality measures for large networks is emerging.

Another limitation for centrality calculations come to the force in dynamic networks [16]. Longitudinal networks have a long tradition [32] [39]. But if we assume that Sampson had the data of 25,000 novices in 500 different but interacting monasteries or Newcomb had millions of observations of thousands of students from the entire campus, both using e-mail or mobile communication logs, the need for efficient algorithms is obvious. Complex calculations are inapplicable for analyzing networks in real time. For instance, if we have to recalculate our measure every minute to monitor transformations in changing networks, measures with, for example, fifteen minutes of calculation time would produce delays in analysis. Another area with the need for an enormous amount of calculations for changing networks can be described as *what-if analysis* where researchers are interested in network change after potential interventions at networks. We ask, how can the intuition provided by well grounded network measures of centrality be preserved in "variations" of those metrics that are scalable and so usable in the massive data context.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

To deal with these challenges some approaches are offered in literature. Borgatti and Everett [10] suggest calculating betweenness centrality using the ego network of every node and show that this a good approximation for betweenness centrality calculated for the overall network. Brandes and Pich [13] use an idea by Eppstein and Wang [22] to show that the shortest path calculation for a selection of nodes (*pivots*) can be used to estimate centralities based on shortest paths. A similar approach is used by Okamato et al. [34] to find the nodes with the highest closeness centrality value in a network. Carpenter et al. [17] recommend decomposing large networks at single node cut points (*articulation points*) or at cuts being more complex, and therefore time consuming, to calculate. Chan et al. [18] offer community centrality for modular networks when the community structure is known or pre-calculated. The advantage of network decomposition lies in the algorithmic complexity of measures based on shortest paths; dividing a network into halves and do the calculation separately for theses halves is much faster than the calculation for the complete network. The dynamic aspects of shortest path algorithms are tackled by Demetrescu and Italiano [20]. They present an algorithm for dynamically updating the shortest path calculation in changing networks without recalculating the overall measure.

In this article we use the concept of calculating the shortest paths from all nodes but restrict the path distances. In the context of betweenness centrality, Borgatti and Everet [11] suggest the term "*k*-betweenness centrality" when limiting the length of the shortest paths. Brandes [14] shows the modification of the algorithm for calculating betweenness centrality [12] needed to calculate *k*-betweenness. Ercsey-Ravasz and Toroczkai [23] showed that the "high-betweenness backbone" of a network can be identified by focusing on sub-graphs created by limited distance paths from nodes. We combine these ideas into a common framework of a *k*-measure approach for closeness and betweenness based centralities. We claim criteria for *k*-measures and elaborate aspects of normalization, implementation, and algorithmic complexity. We show that *k*-centralities are good approximations for the corresponding centrality measure. Therefore, we suggest using these *k*-centrality measures instead of the corresponding centrality measures for larger networks or in cases when the calculation time is critical (dynamics network analysis). In section 2, definitions of graph theory and the basic concept of *k*-centrality measures are introduced. In section 3, we discuss betweenness based measures and in section 4 closeness based measures. These sections also discuss how to modify the algorithms to turn them into *k*-measures algorithms. In section 5, we discuss the impact on the calculation time and in section 6 the validity of the approximations of *k*-centrality measures using random as well as stylized networks. Section 7 consists of case studies using real world networks. The conclusions are discussed in section 8.

2. DEFINITIONS

One of the main reasons for the popularity of centrality measures is their independence from the purpose of the network. Regardless of whether infrastructural networks [44], research cooperation networks [36], or biological networks [19] are being investigated, the same measures within the same software tools are used. This is the case because of the fact that a lot of centrality measures just use the underlying graph structure of networks for calculation. So, from a network analytical perspective, a graph is a network without meaning. Graph theory defines a *graph* $G = (V,E)$ as a set of *nodes* V and a set of *edges* E connecting the nodes. We denote the number of nodes with $N = |V|$ and the number of edges with $M = |E|$. Two nodes n_i and n_j are *adjacent* if and only if an edge e_{ij} exists in E. The number of adjacent nodes d_i for a node n_i is the un-normalized degree of the node. Pendants are nodes with $d_i = 1$. If ω is a weight function of E with $\omega(e) > 0$, $e \in E$ then G is *unweighted* if $\omega(e) = 1$, $e \in E$ and *weighted* otherwise. In *undirected* graphs each edge is reciprocal: $e_{ij} = e_{ji}$, $e \in E$, not so in *directed* graphs.

Two nodes n_i and n_j are *reachable* if there exist a set of nodes $\{n_i, n_a, \ldots, n_f, n_j\} \in V$ and a set of edges $\{e_{ia}, \ldots, e_{fj}\} \in E$. This set of nodes and edges is called a *way*. A *component* is a set of reachable nodes. A way with no nodes or edges occurring more than once is called a *path*. The length of a path $d(n_i, n_j)$ is the number of edges of the path from n_i to n_j. The *shortest path* between a pair of nodes is also known as the *geodesic distance*. The *characteristic path length* is the average of the shortest paths between all pairs of nodes of a network. The longest geodesic distance within a given graph is called *diameter*. The clustering coefficient CC_i for a node n_i is defined as the proportion of actual connections between neighbors of i and all possible connections between the neighbors of i: $|e_{an}|*|e_{an}-1|/2$. The k-reach N_i^k of a node n_i is the number of other nodes which can be reached on all possible paths with the length k starting from n_i. For simplicity we assume unweighted, undirected, and connected graphs for further discussions. Nevertheless, these considerations can be applied to directed or weighted graphs and, with modifications, to unconnected graphs.

Centrality measures [27][43] are used to reveal nodes with certain structural aspects considered as important. The idea of *k*-centralities is based on reducing *global* calculations to *local* calculations. Fig. 1 shows a network with 13 nodes and a star like structure. The central node *13* connects four node chains with three nodes each. Algorithms based on shortest paths (see sections 3 and 4) would calculate the shortest paths from each node to every other node using a breadth-first algorithm [40]. The breadth-first algorithm starts in a given network at any node and *walks* in concentric circles through the network using the edges connecting the nodes. In the first step, all direct neighbors are reached. In the second step, all nodes with distance *2* are reached, etc. So, assuming $k = 2$: Instead of walking from a starting point (e.g. node *1*) via shortest paths through the whole network until reaching nodes *4*, *7*, and *10* the algorithm stops after *2* steps *visiting* nodes *2* and *3*. Starting at node *12* a *2* step limited breadth-first search reaches the nodes *3*, *6*, *9*, *10*, *11*, and *13*.

In the following sections we show some centrality algorithms based on shortest paths and elaborate how to adopt the algorithm of a centrality measure C to become a *k*-centrality measure C_k. We focus on betweenness centrality and closeness centrality and describe other measures just in part when they can be calculated using an almost identical algorithm. For all *k*-centrality measures

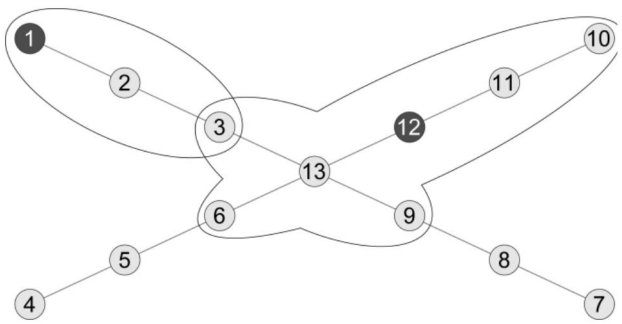

Figure 1: k=2 steps for k-centralities

we define $k \geq 2$, ignoring the case of $k = 1$. We claim three criteria for C_k measures:

a) C_k must remain unchanged the basic ideas of the algorithm of C.

b) C_k has to be an approximation for C.

c) C_k has to converge to C with $C_k = C$, when $k =$ *diameter*.

3. BETWEENNESS BASE MEASURES

The concept of *betweenness* is one of the "three distinct intuitive conceptions of centrality" introduced by Freeman ([27]: 215). Anthonisse [1] defines the *rush* of a node as the total flow through a node on shortest paths. Freeman [27] connects the probability of this in-between position with the notion of information control. These network positions are also responsible for concatenating different parts of a network [15]. Looking at fig. 1, it is obvious that node 13 is important for the *flow* of information through the network starting from any other node.

The betweenness concept is first used by Shimbel [41] to calculate values for every single node in a given network. Shimbel's stress centrality C^S points out that if a node n_a is in an intermediate position between two other nodes n_i and n_j, then n_a has a "certain responsibility" to n_i and n_j. This results in stress if there is activity in the network. C^S calculates the shortest paths for each node to every other node and counts the number of shortest paths including node n_a:

$$C^S(n_a) = \sum_{i<j} g_{ij}(n_a) \qquad (1)$$

Shimbel [41] does not offer any normalization to make the measure comparable for networks with different sizes. However, it is obvious that the center of a star [27] is the position with the highest possible stress. Using C^B_{max} (eq. 2) from betweenness centrality C^B [27] can therefore be used to normalize C^S. This results in a measure similar to betweenness centrality [27] but ignoring the influence of alternative shortest paths to the importance of a specific shortest path. Because C^S is a less known and less used simplification of C^B, we discuss the *k*-measure approach in the following paragraphs using C^B.

Algorithm 1. *k*-centrality for closeness and betweenness based measures; based on [14]

```
for s ∈ V                                              //shortest-k-paths
    empty Predecessors
    dist[] = ∞;  σ[] = 0
    dist[s] = 0; σ[s] = 1; N_s^k = 0
    s → Queue
    while Queue not empty
        Stack ← v ← Queue
        for w such that e_vw ∈ V
            if dist[w] = ∞                                 //new node
                dist[w] = dist[v]+1
                if dist[w] ≤ k
                    w → Queue
                    N_s^k += 1
            if dist[w] = dist[v] + 1
                σ[w] = σ[w] + σ[v]
                v → Predecessors[w]

    while Stack not empty                          //centrality accumulation
        w ← Stack
        for v ∈ Predecessors[w]          //betweenness accumulation
            δ[v] += (σ[v]/σ[w]) · (1 + δ[w])
        if w ≠ s
            C^B[w] += δ[w]
        Distances[s] += dist[w]                    //closeness accumulation

for s ∈ V                                                  //normalizations
    C_k^B[s]' = C_k^B[s] / ((N²−3N+2)/2) · (N(N−1)/(2 Σ_i N_i^k))
    C_k^C[s]' = 1/Distances[s] · (N_s^k/(N−1)) / (1/N_s^k)
```

Freeman [27] points to the independent development of betweenness centrality C^B by Anthonisse [1] and Freeman [26]. C^B is the probability for a node n_a being part of the shortest path connecting two other nodes ([43]: 190). Let g_{ij} be the number of shortest paths from n_i to n_j and $g_{ij}(n_a)$ the number of shortest paths from n_i to n_j including node n_a. The betweenness centrality C^B for a node n_a is consequently

$$C^B(n_a) = \sum_{i<j} \frac{g_{ij}(n_a)}{g_{ij}} \quad with \quad C^B_{max} = \frac{(N^2-3N+2)}{2} \qquad (2)$$

The betweenness score for all nodes in a given network is within the range of *0* (node included in no shortest path) and C^B_{max} (node included in all shortest paths - the center of a star). C^B_{max} is used for normalization to make networks with different sizes comparable.

Based on the introduced idea of *k*-centralities, we limit the path length for the breadth-first search of the shortest paths to *k* steps. In addition, we normalize the scores with the proportion of actual calculated shortest paths $\sum_i N_i^k$ in all possible dyads. The *k*-reach for all nodes $\sum_i N_i^k$ can be easily gathered by counting the number of enqueued nodes during the breath-first searches (see algorithm 1). C_k^B for a node n_a is therefore defined by

$$C_k^B(n_a) = \sum_{i<j, d(n_i,n_j)\leq k} \frac{g_{ij}(n_a)}{g_{ij}} \cdot \frac{N(N-1)}{2\sum_i N_i^k} \qquad (3)$$

Analogous to the definition of C^B, C_k^B is the probability for a node being part of shortest path with the maximum length of k connecting two other nodes. The underlying assumption that C_k^B is an approximation for C^B is that nodes which are between on short paths are more likely to be between on longer paths because short paths are many times part of longer paths as well. The adoption of the algorithm of C^B to calculate C_k^B is simple: stopping the enqueuing process in the breadth-first search algorithm ([14]: 139). Table 1 shows the calculation of *k*-betweenness centrality for the network of fig. 1. The different columns are the results for $k = \{2, 3, 4, 5, 6 = diameter\}$. We can see an underestimation of all centrality scores >0 for small *k*. The results of table 1 suggest that nodes which are in-between on short paths are disproportionately high in-between on longer paths. We do not include this into the normalization of the results because when calculating approximations with *k*-measures, we are more interested in fitting the ranking and the distribution of the result vector rather than the exact values. When we look at the criteria we claimed for *k*-measures at the end of section 2, we see that the *k*-approximation of the final result works fine from a proportional perspective.

4. CLOSENESS BASED MEASURES

Closeness based measures use the concept of shortest paths in a different way. Having short paths to all other nodes is important for efficient communication [7] or information diffusion [6]. The most important closeness based measure is closeness centrality

C^C. The sum of the shortest paths $d(n_a,n_i)$ from a node n_a to all other nodes is the *farness* [37]. The inverse of the farness is the closeness centrality [27]

$$C^C(n_a) = \frac{1}{\sum_{i \neq a} d(n_a,n_i)} \quad with \quad C^C_{max} = \frac{1}{(N-1)} \qquad (4)$$

Therefore, a node is central in case of small shortest path distances to all other nodes. C^C scores again highest for the central node of a star structure (Freeman, 1979) by having a farness of *1* step to all other nodes. To normalize C^C we divide by C^C_{max} [43] to be able to compare networks with different sizes.

The basic assumption of *k*-closeness centrality C^C_k is that nodes which reach a lot of other nodes within *k* steps are more likely to reach all nodes within few steps. Using the idea of bounded-distances on eq. 4 results in the first part of eq. 5. To actually calculate C^C_k we have to additionally consider the size of the *k*-reach N^k_a of a node. Having a lot of nodes within a *k*-neighborhood results in a high value for the sum of the distances. On the other hand, a node with a small N^k_a is treated preferentially. To compensate for this effect, we normalize the distance sum with the proportion of N^k_a within the number of all other nodes (*N-1*). We also adapt $C^C_{k_{max}}$ to the star structure with the size N^k_a. C^C_k for a node n_a is therefore the closeness to all reachable nodes within *k* steps compared to the proportion of reachable nodes after *n* steps:

$$C^C_k(n_a)' = \frac{1}{\sum_{i \neq a, d(n_a,n_i) \leq k} d(n_a,n_i)} \cdot \frac{N^k_a}{N-1} \quad with \quad C^C_{k_{max}}(n_a) = \frac{1}{N^k_a} \qquad (5)$$

The number of nodes reachable for a node n_a within *k* steps can be easily achieved. N^k_a is the number of nodes enqueued during the breadth-first search algorithm (see algorithm 1) when calculating C^C_k. Table 2 shows the results of the approximation steps using the network visualized in fig. 1. The claimed criteria for *k*-centralities fit.

5. PERFORMANCE

In the last sections we derived *k*-centrality measures from their corresponding centrality measures. In this section we analyze the suitability of using *k*-centrality measures alternatively to the centrality measures. Because all measures introduced in the previous sections can be calculated with the same algorithm (algorithm 1), the calculation time for these measures and also its corresponding *k*-measures is the same. Therefore, we do not analyze the performance of a single measure but focus on the running time of the underlying algorithm. We use random and stylized networks for performance and validity evaluation (see next section). Erdos-Renyi random networks [24] are the typical model for random networks. A set of nodes is created and for every pair of nodes a link is set with a constant probability *p*. These networks have short average path lengths and almost no local clustering. Small-World networks [44] are constructed by starting with a *1*-dimensional lattice connecting every node with its *k physically* next neighbors. With a probability *p,* the edges are randomly rewired in the next step, resulting in a network with both a high clustering coefficient and low average distances. Scale-Free networks [3] are created iteratively by adding one node after another and connecting the new nodes with preexisting nodes by favoring nodes which are better connected. This leads to highly centralized networks and *power-law* degree distributions.

Table 3 shows the results of the performance tests for random and stylized networks (Small-World: $p = 0.15$, Scale-Free: $N_0 = 8$). To analyze the impact of the structure of the networks on the calculation time, we have kept the number of nodes (N = 5,000) and the average degree ($\bar{d} = 6.0$) stable. Every line in table 3 reflects the average of the calculations of 100 networks with the same parameters. The time columns show the proportion of the calculation time compared to the original algorithm. The time columns C^x_k in table 3 are therefore a representation of the percentage of the calculation time needed for C^x_2, C^x_3, … (For example, in a Small-World network with $N = 5{,}000$, $\bar{d} = 6.0$, and $p = 0.15$ it takes *4.6* % of the calculation time of a measure to calculate the corresponding *k*-measure with $k = 4$.)

In table 3 one can see the performance advantages of the *k*-measure approach. The locally structured Small-World networks result in the biggest time gain while the very centralized Scale-Free networks perform with the least time gain. The Erdos-Renyi random networks score better than the Scale-Free networks but much worse than Small-World networks. Different network structures result in different time needed to calculate *k*-measures. By taking a closer look at algorithm 1 for calculating the *k*-measures, we are able to see that the breadth-first search algorithm enqueues all neighboring nodes of preselected nodes until the *k*-distance is reached. Therefore the time gain for the *k*-measures is a function of how many nodes are enqueued or checked to be enqueued. If the number of enqueueing attempts within a *k*-distance of all nodes is very high, the performance of *k*-measures is much smaller than in cases when this number is smaller. Looking again at table 3, one can see this effect in the performance values for Scale-Free networks. The centralized structure of these networks and an average path length of *4.0* connect almost every node with the central hubs within 3 steps.

Table 1: *k*-betweenness centrality example

Node(s)	C^B_2	C^B_3	C^B_4	C^B_5	$C^B_6 = C^B$
13	0.273	0.506	0.709	0.788	0.818
3,6,9,12	0.045	0.141	0.217	0.279	0.303
2,5,8,11	0.045	0.056	0.098	0.131	0.167
1,4,7,10	0.000	0.000	0.000	0.000	0.000

Table 2: *k*-closeness centrality example

Node(s)	C^C_2	C^C_3	C^C_4	C^C_5	$C^C_6 = C^C$
13	0.444	0.500	0.500	0.500	0.500
3,6,9,12	0.300	0.355	0.387	0.387	0.387
2,5,8,11	0.188	0.231	0.270	0.300	0.300
1,4,7,10	0.111	0.125	0.167	0.205	0.235

Table 3: Performance in Random and Stylized Networks

Network	Nodes	Avg. Degree	Dia-meter	Path-Length	Time C_2^x	C_3^x	C_4^x	C_5^x
Erdos-Renyi	5,000	6.0	9.3	4.9	0.014	0.061	0.286	0.770
Small-World	5,000	6.0	11.8	7.0	0.008	0.018	0.046	0.126
Scale-Free	5,000	6.0	7.0	4.0	0.022	0.193	0.733	0.999

With the next step, the majority of nodes can be reached, resulting in a rapid increase of calculation time. On the other hand, the high clustering coefficient in Small-World networks [44] implicates a large proportion of the neighbors of a node being connected with each other and therefore bringing a much smaller number of *new* nodes into the enqueueing process. This results in shorter calculation time.

We are able to generalize the estimation of the performance of the *k*-measures for all networks with different structures (including real world networks). The number of enqueueing attempts for a node n_i is its degree d_i. The *k*-distance bounded breadth-first search for a node n_a therefore requires

$$D_a = d_a + d_{x_1} + d_{y_1} + \cdots + d_{x_2} + d_{y_2} + \cdots \quad \text{with } d(n_a, n_{n_k}) = k \qquad (6)$$

enqueueing attempts. We denote $\overline{D}$ as the average value for all D_a in a network. $\overline{D}$ is the key to numerically understanding the performance differences of table 3. The higher the value of $\overline{D}$, the more calculation are processed, and the longer it takes to calculate the *k*-measures. Because the sum of the degree of all nodes in an undirected network is twice the number of edges, *2M,* this is the boundary value of $\overline{D}$ for *k = diameter*. As the time needed to calculate *k*-measures is a function of $\overline{D}$, we can use eq. 7 to estimate the overall amount of time needed for the calculation of a measure C^x after calculating the corresponding *k*-measure C_k^x. The calculation of D_a (and therefore of $\overline{D}$) can be accomplished by calculating d_a for all nodes before starting the *k*-measure algorithm and summing up these values during the enqueueing process of every node without affecting the calculation complexity.

$$t_{C_k^x} = t_{C^x} \cdot \overline{D} \cdot \frac{1}{2M} \quad with \quad \overline{D} = \frac{\sum_{n_a}(D_a)}{N} \qquad (7)$$

Applying these considerations to networks with different sizes but stable average degree (e.g. interpersonal communication networks [28]), one important conclusion can be drawn: The time needed to calculate *k*-distance bounded breadth-first search for a single node n_a is independent from the network size and is solely a function of the average local structure $\overline{D}$ dominated by the network topology and the average degree[1]. This implies another advantage of *k*-measures for large networks. The time needed for calculation can be described with $\Theta(n\overline{D})$. As $\overline{D}$ is a constant value independent from the network size, the calculation complexity of *k*-centrality measures is $\Theta(n)$. This is a significant improvement over the *Θ(nm)* of the original algorithm [12].

[1] For example, for the stylized test networks $\overline{D}$ = 131 for *k* = 2 and $\overline{D}$ = 421 for *k* = 3 for small world networks; $\overline{D}$ = 291 for *k* = 2 and $\overline{D}$ = 1,726 for *k* = 3 for Erdos-Renyi networks.

6. VALIDITY

Table 4 shows the validity test for C_k^B compared to $C^B$with *k=2, k=3, k=4* and *k=5.* The validity is measured in three ways. First, we correlate (Pearson) the C_k^Bresults with the C^Bresult. Because calculations of betweenness centrality do not produce normally distributed result vectors and the independence assumption for correlations is violated with network data, we use two additional performance indicators for validation of the results. Top 10 Hits: The number of the top 10 nodes of the calculated measures which are correctly identified using *k*-calculations. Top 1: The *k*-distance which is (on average) needed to correctly identify the most important node. Examine the validity of identifying the most central nodes is important because this is the essential task of centrality measures, In table 4 you can see that in Small-World networks with *5,000* nodes and an average degree of *6.0,* using *k = 4* results in the ability to identify *7.9* top ten nodes correctly and also allows the most central node to be found. Table 5 shows the same performance and validity experiments for C_k^C with similar outcome. Both tables are based on average values for *100* networks with identical parameters.

In tables 4 and 5, one can see the connection between network structure and validity. Shorter average path lengths in random and scale-free networks bring every node in near path distance to the center(s) of the network. Therefore the globally important nodes can be identified with smaller k. On the other hand, the high clustering coefficient in small-world networks results in a flatter slope of calculation time (see previous section) but also in less validity.

7. CASE STUDIES

In the previous section we discussed issues of performance and validity using random and stylized networks. In this chapter we apply the ideas of *k*-centralities to real world data. The first network is a medium sized collaboration network. We use this network to analyze performance and validity. The second network is a large network, and we therefore do not calculate the centralities but the *k*-centralities and use this to estimate the calculation time for the corresponding centrality measures.

7.1 Astro-Ph: A Collaboration Network

Newman [33] created a network of scientists posting preprints on the "High-Energy Theory E-Print Archive" between Jan 1, 1995 and December 31, 1999. Two scientists are connected if they are co-authors of at least one common paper. We ignore the line weights in the network and focus on the largest component, including 14,842 nodes and 119,616 edges (avg. degree 16.1). Newman describes the network as having *small-world* attributes (avg. path length 4.8, diameter 14). Therefore, we expect good performance of *k*-centralities. On the other hand the high, average degree will affect the performance adversely. The size of the network allows us to calculate the centrality measures and compare the results with the *k*-centrality measures.

Table 4: Validity of *k*-betweenness centrality applied to random and stylized networks with 5,000 nodes and an average degree of 6.0

Network	Dia-meter	Path-Length	Pearson Correlation C_2^B	C_3^B	C_4^B	C_5^B	Top 10 Hits C_2^B	C_3^B	C_4^B	C_5^B	Top 1 C_n^B
Erdos-Renyi	9.3	4.9	0.978	0.992	0.990	0.995	6.9	9.0	8.9	9.4	2.9
Small-Worlds	11.8	7.0	0.919	0.974	0.990	0.991	5.2	6.6	7.9	8.2	3.5
Scale-Free	7.0	4.0	0.975	0.990	0.999	1.000	9.3	9.7	9.9	10.0	2.2

Table 5: Validity of *k*-closeness centrality applied to random and stylized networks with 5,000 nodes and an average degree of 6.0

Network	Dia-meter	Path-Length	Pearson Correlation C_2^C	C_3^C	C_4^C	C_5^C	Top 10 Hits C_2^C	C_3^C	C_4^C	C_5^C	Top 1 C_n^C
Erdos-Renyi	9.3	4.9	0.940	0.958	0.975	0.964	9.1	9.6	9.6	9.9	2.2
Small-Worlds	11.8	7.0	0.908	0.941	0.954	0.966	6.3	7.8	9.0	9.5	2.9
Scale-Free	7.0	4.0	0.974	0.968	0.996	1.000	9.9	10.0	10.0	10.0	2.0

Table 6 shows the results of the performance and validity calculations for the *astro-ph* network for *k*-betweenness and *k*-closeness centrality. The short average path length, a result of a higher degree and a small-world structure, are mentioned by Newman (2001). These attributes describe the big increase in calculation time from C_2^x to C_3^x. The most central node can be identified with $k = 2$ while some nodes in the top *10* betweenness centrality ranking are incorrect. This results in a weaker validity for betweenness centrality.

Table 6: Performance and Validity for astro-ph network

	Time C_2^x	C_3^x	Correlation C_2^x	C_3^x	Top 10 Hits C_2^x	C_3^x	Top 1 C_n^x
Betweenness	0.088	0.342	0.820	0.876	7	8	2
Closeness	0.088	0.342	0.903	0.981	10	10	2

7.2 Wikipedia: A Large Network

The network consists of 659,388 nodes representing English Wikipedia pages and 1,182,967 un-directed edges linking the pages with each other in a case of bi-directed links. The average degree is 3.6, the maximum degree is 1,488. The network is based on the Wikipedia XML corpus collected by Denoyer and Gallinari [21] and was created for the 2007 Sunbelt Viszards-Session [4]. To approximate the most central nodes in the Wikipedia network, we use $k=2$ and calculate *k*-betweenness C_2^B and *k*-closeness centrality C_2^C. The calculation takes 33 minutes on a personal computer with an Intel Xeon 2.80 GHz CPU and 12 GB Ram without the use of multi processors. We use eq. 3 to estimate a calculation time for the centrality measures C^x of 25 days with the same system. Of course, it is possible to do these calculations with the help of super computer or computer clusters using parallel algorithms [2]. However, the Wikipedia network is too large to calculate betweenness and closeness centrality on a single home computer within an acceptable period of time. By using *k*-centrality measures we are able to approximate measures based on shortest paths within reasonable time.

8. DISCUSSION

Borgatti and Everett [2006] suggested calculating betweenness centrality using bounded-distance shortest paths calculations. Brandes [14] offered an efficient algorithm for this calculation. In this article we elaborated the idea of using *k*-centrality measures as approximations for closeness and betweenness based centrality measure. We claimed criteria for *k*-centrality measures, offered normalizations, and analyzed performance and validity. In a nutshell, *k*-centrality measures bring a tremendous gain of calculation time and also a high validity of the results. We were able to show that just 5 % of the original calculation effort is needed to receive a valid outcome. In denser or highly centralized (e.g. scale-free) networks, shorter average path distances prevent the time benefits of *k*-centrality measures. If the average degree is low (e.g. 5 to 25) and the network is locally structured (e.g. small-world networks), the time gain is enormous. A small average degree and high local clustering are essential attributes of networks constructed of human interactions [28]. Therefore, it is possible to calculate *k*-centrality measures based on shortest paths for such networks with a few hundred of thousands or even millions of nodes within feasible time. We were also able to show that the calculation complexity of *k*-centrality measures is $\Theta(n)$ in case the average degree is constant with increasing network size, which results in a much better scalability than the most efficient algorithm for betweenness and closeness centrality [12] with $\Theta(nm)$.

The reduction of the calculations leads to a massive improvement in scalability. This is not surprising. More interesting is the high validity of the results by just using a small fraction of the original calculations. We explain this outcome by considering the nature of shortest paths in typical networks. First, it is clear by definition that the less important a node is, the larger its distance to all other nodes. Networks are structured in a way such that paths into the center are short. This is true for random and stylized networks (e.g. small-world, scale-free networks) as well as for real world networks. Therefore, almost every node is connected to and through the center on short paths. However, only the nodes in the center have small distances to all the peripheral nodes. Therefore, using only *k* steps for the calculation of shortest path for every node means that the algorithm focuses mostly on central nodes and rarely on peripheral nodes. Second, the majority of the nodes do not contribute to the overall structure of a network; particularly in the core-periphery networks common within communities, organizations and task-groups. In such networks, algorithms which calculate the shortest paths between all nodes spend more of their time focusing on peripheral nodes. Preventing the breadth-first search algorithm to *walk* to every single peripheral pendant during the shortest path calculations for all nodes thus saves substantial computational time with minimal impact on the magnitude of the resulting metric.

This argument suggests a drawback to *k*-centrality measures: The incapability of detecting global bridges when they are distant from denser connected centers. When we assume a network designed of two almost separated components of actors, both very well connected within the groups and simply connected by a chain of the length of, for example, five actors with each other. The actor in the middle of the connection chain would score highest in betweenness and also in closeness centrality. *k*-centrality measures are not able to identify this actor (unless *k* is set unprofitably high). Thus, there is also an argument for local calculations with regards to the contents of larger social networks consisting of human beings. Normally information does not pass over many steps to every single node in the complete network. In real life the assumed actor in the middle of a connecting chain might sit on the fence rather than in the center.

In this article we described local algorithms for the approximation of centrality measures based on shortest paths. These are the most widely-used centrality algorithms. A key issue for future research is whether measures based on shortest paths are really appropriate for all networks or only for certain classes of networks. Another focus of subsequent research has to tackle the question of how to estimate the *k* of the *k*-measures to guarantee satisfying results. We do not address these issues here; however, we note that in some networks, the rate of the "item" being sent through the network (e.g. information, beliefs or diseases) may atrophy or decay at such a rate that a *k*-centrality measure may be a better measure of node criticality. Future research should consider this issue.

In addition, there are other measure families, such as measures using all paths [42] or eigenvector based measures [8][29]. Future work should consider the implications of the *k*-centrality measures approach as approximations of the corresponding centralities for these measure families.

9. ACKNOWLEDGMENTS

This work is supported in part by the Office of Naval Research (ONR), United States Navy (ONR MURI N000140811186, ONR MMT N00014060104). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research or the U.S. government. We are grateful to Ulrik Brandes for his insightful comments. We are thankful to Harald Katzmair and Max Ruhri for discussing first ideas which led to this article.

10. REFERENCES

[1] Anthonisse, J. M. 1971. *The rush in a directed graph*, Technical Report BN 9/71, Stichting Mathematisch Centrum, Amsterdam, Netherlands.

[2] Bader, D. A., and Madduri, K. 2006. Parallel algorithms for evaluating centrality indices in real-world networks. *Proceedings of the ICPP 2006 - International Conference on Parallel Processing* , 539-550.

[3] Barabási, A. L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286, 509 - 512.

[4] Batagelji, V., Börner, K., Brandes, U., Hong, S.H., Johnson, J., Krempel, L., Mrvar, A., Pfeffer, J., and Quan, W. 2007. Visualization and analysis of Wikipedia. *Viszards-Session at the Sunbelt XXVII Conference*.

[5] Batagelj, V., Kejžar, N., Korenjak-Cerne, S., and Zaveršnik, M. 2006. Analyzing the structure of U.S. patents network. *Data science and classification*, eds. V. Batagelj, H.-H. Bock, A. Ferligoj, A. Iberna, Springer, Berlin.

[6] Bavelas, A. 1948. A mathematical model for group structure. *Human Organization* 7, 16-30.

[7] Beauchamp, M. A. 1965. An improved index of centrality. *Behavioral Science* 10, 161-163.

[8] Bonacich, P. 1972. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology* 2, 113-120.

[9] Borgatti, S. P., Everett, M.G., and Freeman, L.C. 2002. *Ucinet 6 for Windows*, Analytic Technologies, Cambridge, MA, USA.

[10] Borgatti, S. P. 2005. Centrality and network flow. *Social Networks* 27, 55-71.

[11] Borgatti, S. P., and Everett, M.G. 2006. A graph-theoretic perspective on centrality. *Social Networks* 28, 466-484.

[12] Brandes, U. 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology* 25, 2, 163-177.

[13] Brandes, U., and Pich, C. 2007. Centrality estimation in large networks. *International Journal of Bifurcation and Chaos* 17, 7, 2303-2318.

[14] Brandes, U. 2008. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks* 30, 136-145.

[15] Burt, R. S. 2005. *Brokerage and Closure: An introduction to social capital*, Oxford University Press, New York, NY, USA.

[16] Carley, K. M. 1999. On the evolution of social and organizational networks. *Special issue of Research in the Sociology of Organizations*, eds. S.B. Andrews, D. Knoke, JAI Press, Greenwhich, CT, USA.

[17] Carpenter, T., Karakostas, G., and Shallcross, D. 2002. Practical issues and algorithms for analyzing terrorist networks. *Invited Paper: Western Multi-Conference (WMC 2002)*..

[18] Chan, S. Y., Leung, I.X.Y., and Liò, P. 2009. Fast centrality approximation in modular networks. Proceeding of the 1st ACM international workshop on Complex networks meet information & knowledge management (CNIKM '09) , 31-38.

[19] Chase, I. D. 1982. Dynamics of hierarchy formation: The sequential development of dominance relationships. *Behaviour* 80, 218-240.

[20] Demetrescu, C., and Italiano, G. 2001. Fully dynamic all pairs shortest paths with real edge weights. *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science* , 260-267.

[21] Denoyer, L., and Gallinari, P. 2006. *The Wikipedia XML corpus*, http://www-connex.lip6.fr/ ~denoyer/Wikipedia XML/ [6/2/2011].

[22] Eppstein, D., and Wang, J. 2004. Fast approximation of centrality. *Journal of Graph Algorithms and Applications* 8, 1, 39-45.

[23] Ercsey-Ravasz, M., and Toroczkai, Z. 2010. Centrality scaling in large networks. *Phys. Rev. Lett.* 105, 038701.

[24] Erdos, P., and Renyi, A. 1959. On random graphs. *Publ. Math., Debrecen* 6, 290-297.

[25] Everett, M. G., and Borgatti, S. P. 2005. Ego network betweenness. *Social Networks* 27, 31-38.

[26] Freeman, L. C. 1977. A set of measures of centrality based on betweenness. *Sociometry* 40, 35-41.

[27] Freeman, L. C. 1979. Centrality in social networks: Conceptual clarification. *Social Networks* 1, 215-239.

[28] Hamill, L., and Gilbert, N. 2009. Social Circles: A Simple Structure for Agent-Based Social Network Models. *Journal of Artificial Societies and Social Simulation* 12, 2.

[29] Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. *ACM* 46, 604-632.

[30] Leavitt, H. J. 1951. Some effects of certain communication patterns on group performance. *Journal of Abnormal and Social Psychology* 46, 38-50.

[31] Leskovec, J., Huttenlocher, D., and Kleinberg, J. 2010. Governance in social media: A case study of the Wikipedia promotion process. *AAAI International Conference on Weblogs and Social Media (ICWSM '10)*.

[32] Newcomb, T. N. 1961. *The acquaintance process*, Holt, Rinehart and Winston, New York, NY, USA.

[33] Newman, M. E. J. 2001. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* 98, 404-409.

[34] Okamoto, K., Chen, W., and Li, X.-Y. 2008. Ranking of Closeness Centrality for Large-Scale Social Networks. *Proceedings of the 2nd annual international workshop on Frontiers in Algorithmics (FAW '08)*, 186-195.

[35] Onnela, J. P., Saramäki, J., Hyvönen, J., Szabó, G., de Menezes, M. A., Kaski, K., Barabási, A. L., and Kertész, J. 2007. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics* 9.

[36] Palinkas, L. A., Johnson, J. C., Boster, J. S., and Houseal, M. 1998. Longitudinal studies of behavior and performance during a winter at the south pole. *Aviation, Space, and Environmental Medicine* 179, 73-77.

[37] Sabidussi, G. 1966. The centrality index of a graph. *Psychometrika* 69, 581-603.

[38] Sade, D. S. 1989. Sociometrics of macaca mulatta III: N-path centrality in grooming networks. *Social Networks* 31, 273-292.

[39] Sampson, S. F. 1968. A novitiate in a period of change. An experimental and case study of social relationships, PhD Thesis, Cornell University, NY, USA.

[40] Sedgewick, R. 2011. *Algorithms*. 4th Edition, Addison-Wesley, Reading, MA, USA.

[41] Shimbel, A. 1953. Structural parameters of communication networks. *Bulletin of Mathematical Biophysics* , 501-507.

[42] Stephenson, K. A. and Zelen, M. 1989. Rethinking centrality: Methods and examples. *Social Networks 11*, 1-37.

[43] Wasserman, S., and Faust, K. 1995. *Social Network Analysis, Methods and Applications*, Cambridge University Press, Cambridge, UK.

[44] Watts, D., and Strogatz, S. 1998. Collective dynamics of small world networks. *Nature* 393, 440-442.

[45] Yang, J., and Leskovec., J. 2011. Temporal variation in online media. *ACM International Conference on Web Search and Data Mining (WSDM '11)* 15.