Radboud University Nijmegen

Master Thesis

Local Approximation of Centrality Measures

Author:	Max Hinne, MSc
Supervisor:	Prof. Dr. Ir. Th. P. van der Weide
Second corrector:	Dr. E. Marchiori

Thesis number: 645
January 2011, Nijmegen, The Netherlands

Submitted in partial fulfillment of the requirement of Master of Science in Computer Science at the Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands.

Local Approximation of Centrality Measures

Max Hinne Theo van der Weide

January 16, 2011

Abstract

Centrality measures provide a means to differentiate the importance of vertices in a network. These measures are mathematically clear, but the algorithms to compute them often have quadratic time complexity or worse. This may lead to significant computational challenges when applied to large networks. In this paper, we propose a *local* strategy for three frequently used centrality measures: (i) closeness, (ii) betweenness and (iii) PageRank. This local approach uses only the vertices directly adjacent to a target vertex to derive an approximation of the true centrality measure. The approximations are accompanied with an analysis of the approximation error bounds. Our analysis and experiments show that local approximations are quite successful on undirected graphs, and on directed graphs depending on the reciprocity of edges.

1 Introduction

Many collections of data can be represented as complex networks, such as the world wide web, various social networks, networks of interacting proteins or internet routers. An important aspect of complex and/or social network analysis is formed by centrality indices, measures that are used to indicate the (relative) importance of vertices and edges in the networks [1]. Mathematically, these measures are usually well defined and fairly simple to grasp. However, their actual calculation is much more involved, for the following reasons:

- First of all, the networks we study are exceedingly large. Although most path-based centrality indices can be calculated in polynomial time, running time complexities of $O(n^2)$ or $O(n^3)$ are still prohibitive for networks consisting of millions or even billions of nodes. Specialized algorithms are sometimes able to lower this complexity, but even they are cumbersome for networks of web magnitude.
- Second, the networks are updated continuously, adding or removing vertices and adding, removing or rewiring edges. Being able to track the changes in a network and the accompanying centrality rankings is an important topic, but infeasible with slow algorithms.

The aforementioned issues force us to rely on approximation algorithms instead of calculating the exact centralities. But not any approximation will do. The most straightforward solution to the problem of scale would be to find a representative sample and use this to find an estimated mean and variance of the true centralities [2, 3, 4]. However, to find such a sample, the entire network must be accessible and known which is not a reasonable assumption.

Fortunately, all is not lost. We propose to approximate centrality indices with what we refer to as *local* techniques [5]; algorithms that at each step of their execution consider only vertices adjacent to those they have seen before. For example, such an algorithm could estimate the centrality of a vertex by only examining it, its neighbours and its neighbour's neighbours. With these local approximation techniques, we overcome both problems.

The paper is structured as follows: we take off by introducing the notation and definitions that are used throughout this paper in Section 2. This is followed by Section 3 in which we describe the application of our approach on three well-known centrality measures. Here we also

analyze the approximation errors that accompany our approximations. The performance of the local approximations in terms of running time complexity is discussed in Section 4. Afterwards, we show a series of experiments in Section 5 to validate our theoretical results. Finally, we draw the conclusions of our work in Section 6.

2 Preliminaries

We model a network as a graph $G = (V, E)$, where V_G is the set of vertices and $E_G \subseteq V_G \times V_G$ is the set of edges that connect the vertices. We define a H as a subgraph of G, denoted $H \subset G$, iff its vertices form a subset of the vertices of G, i.e. $H \subset G \triangleq V_H \subset V_G$.

Let n_G be the number of vertices in a graph: $n_G = |V|$ and m_G the number of edges: $m_G = |E|$. The subscript G is omitted when there is no confusion likely to occur. We use the shorthand $v \to w$ to indicate v is connected to w and its transitive closure $v \xrightarrow{+}_G w$ to denote that v is connected to w through a path of one or more edges in G. The shortest path between two vertices is called a geodesic path, its length is the distance $d(v, w)$ between its endpoints. We also use the shorthand $v \xrightarrow{k} w \triangleq d(v, w) = k$. Note that in the case of undirected graphs $v \to w$ and $w \to v$ are equivalent (idem for paths), but this not the case in general when the edges are directed.

The neighbourhood $N(v)$ of a vertex v consists of those vertices that it is connected with: $N(v) = \{w \in V \mid v \to w\}$. In the case of directed networks we will make the distinction between the neighbours that connect to v, $N_{\text{in}}(v)$ and those that v connects to, $N_{\text{out}}(v)$. The degree $k(v)$ of a vertex is simply the number of neighbours it has. In-degree and out-degree of v in directed networks is defined analogous with its neighbourhood.

2.1 Definitions

We are looking for the smallest possible subgraph around v on which we can calculate a successful approximation of a centrality measure. For each vertex for which we calculate its centrality, the subgraph for that particular calculation must be local with respect to this vertex. The concept of a local subgraph will be formally defined as:

Definition 2.1

We call a graph H a *local subgraph with respect to a vertex* $v \in V_G$, denoted $\mathcal{L}(G, H, v)$, if it satisfies the following three axioms:

AXIOM 1. $H \subset G$

AXIOM 2. $v \in V_H$

AXIOM 3. $w \in V_H \Rightarrow v \xrightarrow{*}_H w$

In other words, a subgraph is a local subgraph w.r.t. v iff all vertices in the subgraph can be reached from v. In directed networks this implies that a member w of the in-neighbors of v, $w \in N_{\text{in}}(v)$, is not part of the subgraph unless there is a path $v \xrightarrow{+} w$ as well. However, many centrality measures implicitly use such incoming edges in their definitions. To analyze this problem we introduce the concept of *reciprocity-at-distance-k* ρ_k, which we define as:

Definition 2.2

The *reciprocity-at-distance-k* ρ_k of a graph G is the conditional probability that an edge from v to w exists, given that there is a path from w to v of k steps:

$$\rho_k(G) = \mathsf{Pr}\left(w \to v \,\middle|\, v \xrightarrow{k} w\right) \quad . \tag{2.1}$$

Definition 2.3

To obtain a local subgraph, we make use of an *environment generator* g. It assigns to a vertex a subgraph that is strictly smaller than the entire graph. A function is an environment generator, denoted $\mathcal{E}(g)$, iff

1. for each graph G the function $g(G)$ assigns to each node $v \in V_G$ a local subgraph: $\mathcal{L}(G, g(G)(v), v)$ and
2. this local subgraph is essentially smaller than the graph itself. The latter is formalized as follows. Let $\hat{n}_{g(G)}$ be the maximum of $n_{g(G)(v)}$ for all nodes $v \in V_G$. Then the size requirement is expressed as $\hat{n}_{g(G)} = o(n_G)$ for $n_G \to \infty$.

The requirement that the assigned subgraph must be significantly smaller than the entire graph leads to the critical trade-off in local approximations: smaller subgraphs lead to faster computation, but decreased approximation accuracy.

We consider the functions we intend to approximate – centrality measures – as a general class of functions operating on networks:

Definition 2.4

A *network function* is a function that, given a graph G and a vertex $v \in V_G$, assigns a value to v. Formally, it is a function with signature $f : \mathcal{G} \to V_G \to \mathbb{R}$.

Critically, a network function can be restricted to a specific input domain. This forms the basis of our approach for local approximations:

Definition 2.5

Given an environment generator $g(G)$, a *local network function* with respect to a vertex $v \in V_G$ is a function that, given a graph G, assigns a value to that vertex based on its local subgraph assigned by $g(G)$. Formally, it is a function $f(G)$ with signature $f(G) : V_G \times \wp(G) \to \mathbb{R}$.

Calculations on $g(G)(v)$ will likely differ from the 'true' result on G. We require that this difference is as small as possible. We distinguish between *local approximations* and *local order approximations*:

Definition 2.6

Let f be a network function and G a graph. We call a local network function $\tilde{f}(G)$ a *local approximation* of f if there exists an environment generator that assigns a local subgraph to each vertex v, $g(G)(v)$, such that the computation by $\tilde{f}(G)$ on the subgraph yields a close approximation of the value computed by f. Formally, $\tilde{f}(G)$ is a local approximation iff

$$\forall_{G \in \mathcal{G}} \forall_{v \in V_G} \exists_{g:\mathcal{E}(g)} \left[\left| f(G)(v) - \tilde{f}(G)(v, g(G)(v)) \right| \leq \epsilon \right] \quad . \tag{2.2}$$

From an application perspective, the exact centrality values are not always relevant. What is more relevant is that the order that is induced by a centrality measure stays the same in the approximation.

Definition 2.7

Let f be a network function. Then for any graph G, f induces a *network ordering* $\preccurlyeq_G$ of the vertices V_G as follows:

$$\forall_{v,w \in V_G} \left[v \preccurlyeq_G w \triangleq f(G)(v) \leq f(G)(w) \right] \quad . \tag{2.3}$$

$(V_G, \preccurlyeq_G)$ is a partial order, as it is reflexive, antisymmetric and transitive.

Definition 2.8

Let $\tilde{f}$ be a local network function and g an environment generator. Then for any graph G, $\tilde{f}$ induces a *local network ordering* $\tilde{\preccurlyeq}_G$ of the vertices V_G. Formally:

$$\forall_{v,w \in V_G} \left[v \tilde{\preccurlyeq}_G w \triangleq \tilde{f}(G)(v, g(G)(v)) \leq \tilde{f}(G)(w, g(G)(w)) \right] \quad . \tag{2.4}$$

$(V_G, \tilde{\preccurlyeq}_G)$ is once again a partial order, as it is reflexive, antisymmetric and transitive.

Definition 2.9

Let $\preccurlyeq_G$ be the ordering induced by a network function f on the vertices V_G and let $\tilde{\preccurlyeq}_G$ be a local network ordering induced by $\tilde{f}$. We say that $\tilde{f}$ is a *local order approximation* of f iff

$$\tau(\preccurlyeq_G, \tilde{\preccurlyeq}_G) \geq \delta \ , \tag{2.5}$$

with $\tau \in [0, 1]$ a statistic measuring the correlation between the two orderings.

Although the choice for a specific correlation statistic is in principle arbitrary, in this paper we opt for Kendall's tau-b [6], as it does not pose any restrictions on the distributions of the orderings and makes adjustments for ties.

3 Closeness, betweenness and PageRank

Using the framework of definitions from the previous section, we proceed to describe local approximations for a series of centrality indices. We consider (i) *closeness*, (ii) *betweenness* and (iii) *PageRank*. The first two are widely adopted measures in social network analysis [1]. The latter measure [7] has formed the basis for the success of Google, but is now also used in entirely different contexts, e.g. [8].

For each of these measures we suggest an intuitive local approximation $\tilde{f}$, based on the local connectivity g of a vertex. We then pose two questions for each measure:

Research question 1

Given a centrality measure f, a local network function $\tilde{f}$ and an environment generator g, what is the approximation error

$$\epsilon = \left| f(G)(v) - \tilde{f}(G)(v, g(G)(v)) \right| \tag{3.1}$$

that $\tilde{f}$ induces, in terms of the parameters of g?

And:

Research question 2

Given a centrality measure f, a local network function $\tilde{f}$ and an environment generator g, (to what extend) is $\tilde{f}$ a local order approximation?

We approach the first question in a theoretical fashion, the second question will be answered in the form of a series of experiments.

3.1 Closeness centrality

The *closeness centrality* c_C of a vertex v is given by:

$$c_C(G)(v) = \sum_{w \in V_G} 2^{-d(v,w)} \ . \tag{3.2}$$

It is a measure of the (reciprocal of the) distance between a vertex and all other vertices in the graph. Throughout this analysis we will use the normalized version of closeness centrality:

$$c_C(G)(v) = \frac{1}{n_G} \sum_{w \in V_G} 2^{-d(v,w)} \ . \tag{3.3}$$

Another formulation of closeness centrality allows us to easier distinguish the contributions by vertices at different distances. Let $N_k(v)$ be the vertices that are k edges away from v, i.e.

$N_k(v) = \{w \in V \mid d(v,w) = k\}$. Note that $N_1(v)$ coincides with $N(v)$ and $N_0(v) = \{v\}$. Using this notation, the closeness centrality of a vertex can be expressed as

$$c_C(G)(v) = \frac{1}{n_G} \sum_{k=0}^{\infty} |N_k(v)|\, 2^{-k} \ . \tag{3.4}$$

We create an intuitive local approximation of c_C by choosing the local subgraph as the graph induced by the vertices at most d edges away from v. Let $H(v)$ be this subgraph, then formally $H(v)$ is defined as

$$H(v) = \bigcup_{k=0}^{d} N_k(v) \ . \tag{3.5}$$

Lemma 3.1 $H(v)$ is a local subgraph.

Proof:
$H(v)$ satisfies the three axioms as per Definition 2.1:

1. It follows from the definition that $V_H \subset V_G$ and therefore $H \subset G$.
2. $d \geq 0$, so $N_0(v) \subseteq H(v)$. Since $N_0(v) = \{v\}$, it follows that $v \in H(v)$.
3. If $v, w \in H$, then there is a shortest path between v and w with length at most d. From this it follows directly that $v \xrightarrow{*} w$.

□

Closeness centrality consists of the sum of contributions by vertices in $H(v)$ and those in $G \setminus H(v)$:

$$\begin{aligned} c_C(G)(v) &= \frac{1}{n_G} \sum_{w \in V_H} 2^{-d(v,w)} + \frac{1}{n_G} \sum_{w \in V_G \setminus V_H} 2^{-d(v,w)} \\ &= \frac{1}{n_G} \sum_{k=0}^{d} |N_k(v)|\, 2^{-k} + \frac{1}{n_G} \sum_{k=d+1}^{\infty} |N_k(v)|\, 2^{-k} \ . \end{aligned} \tag{3.6}$$

By considering only the first component, we obtain the approximation $\tilde{c}_C$:

$$\tilde{c}_C(G)(v, H) = \frac{1}{n_G} \sum_{w \in V_H} 2^{-d(v,w)} = \frac{1}{n_G} \sum_{k=0}^{d} |N_k(v)|\, 2^{-k} \ . \tag{3.7}$$

Lemma 3.2 $\tilde{c}_C(G)(v, H)$ has approximation error $\epsilon \leq 2^{-d}$.

Proof:
The second term in (3.6) is the approximation error ϵ of (3.7), for which we can derive

$$\begin{aligned} \epsilon &= \frac{1}{n_G} \sum_{k=d+1}^{\infty} |N_k(v)|\, 2^{-k} \\ &\leq 2^{-d} \ , \end{aligned} \tag{3.8}$$

where we made use of the fact that $|N_k(v)| \leq n_G - n_H \leq n_G$. □

Lemma 3.3 Let $\mathcal{H}_d$ be the function that assigns for each graph G to each vertex the local subgraph as described above. In addition, let D be the longest shortest path from the most central node $\hat{v}$ in G. Then $\mathcal{H}_d$ is an environment generator iff $d < D$.

Proof:
Since the distance between the source vertex v and any vertex in H_d is smaller than D, there are at least some vertices unreachable from v. Yet, H_d is a local subgraph, so $\mathcal{H}_d$ is an environment generator. □

This allows us to conclude:

Lemma 3.4 Let $\mathcal{H}_d$ be the environment generator as described above. Then $\tilde{c}_C$ is a local approximation of c_C with approximation error 2^{-d}.

Proof:
This follows directly by choosing ϵ in Definition 2.2 to be 2^{-d}. □

It is worth noting that the approximation error decreases exponentially with the distance of considered neighbours. This implies that the approximation error is small even when only a small local subgraph is considered. Consequently, closeness centrality is well suited for local approximation.

In Section 5 we show that $\tilde{c}_C$ is also a local order approximation.

3.2 Betweenness centrality

Whereas closeness is an indicator of the number of vertices that can be reached from v, betweenness is a measure of amount of communication that passes through v. It is expressed as the fraction of shortest paths through v. Let $\sigma(u, w)$ be the number of shortest paths between u and w, and let $\sigma(u, v, w)$ be the number of shortest paths between u and w that pass through v. The *betweenness centrality* c_B is then given by:

$$c_B(G)(v) = \sum_{u,w \in V_G \setminus \{v\}} \frac{\sigma(u, v, w)}{\sigma(u, w)} \ . \tag{3.9}$$

Similar to the analysis of closeness, we use the normalized version of this measure instead:

$$c_B(G)(v) = \frac{1}{(n_G - 1)(n_G - 2)} \sum_{u,w \in V_G \setminus \{v\}} \frac{\sigma(u, v, w)}{\sigma(u, w)} \ . \tag{3.10}$$

To obtain a local approximation for betweenness centrality, let us examine the summation terms in isolation. The enumerator in Eq. (3.10) can be decomposed as

$$\sigma(u, v, w) = \sigma(u, v) \cdot \sigma(v, w) \ . \tag{3.11}$$

By repeatedly applying this identity, we may partition the summation terms in (3.10) into those fractions that include a predecessor and a successor of v:

$$\frac{\sigma(u, v, w)}{\sigma(u, w)} = \sum_{\substack{v_0 \to v \to v_1 \\ v_0 \neq v_1}} \frac{\sigma(u, v_0, w)}{\sigma(u, w)} \cdot \frac{\sigma(v_0, v, v_1)}{\sigma(v_0, v_1)} \cdot \frac{\sigma(u, v_1, w)}{\sigma(u, w)} \ . \tag{3.12}$$

Let $S = \mathbb{E}[\sigma(\cdot, \cdot)]$ be the expected number of paths between any pair of vertices. Note that $\sigma(u, v_0, w) \cdot \sigma(u, v_1, w) \leq S$. This implies

$$\begin{aligned} \frac{\sigma(u, v, w)}{\sigma(u, w)} &\approx \sum_{\substack{v_0 \to v \to v_1 \\ v_0 \neq v_1}} \frac{S}{S^2} \cdot \frac{\sigma(v_0, v, v_1)}{\sigma(v_0, v_1)} \\ &= \frac{1}{S} \sum_{\substack{v_0 \to v \to v_1 \\ v_0 \neq v_1}} \frac{\sigma(v_0, v, v_1)}{\sigma(v_0, v_1)} \\ &\propto \sum_{\substack{v_0 \to v \to v_1 \\ v_0 \neq v_1}} \frac{\sigma(v_0, v, v_1)}{\sigma(v_0, v_1)} \ , \end{aligned} \tag{3.13}$$

where we made use of the fact that S is independent of v. Substituting this result into the definition (3.10) yields:

$$c_B(G)(v) \propto \frac{1}{(n_G-1)(n_G-2)} \sum_{u,w \in V_G \setminus \{v\}} \sum_{\substack{v_0 \to v \to v_1 \\ v_0 \neq v_1}} \frac{\sigma(v_0,v,v_1)}{\sigma(v_0,v_1)}$$

$$= \sum_{\substack{v_0 \to v \to v_1 \\ v_0 \neq v_1}} \frac{\sigma(v_0,v,v_1)}{\sigma(v_0,v_1)} \tag{3.14}$$

$$= \tilde{c}_B(G)(v,H) \ , \tag{3.15}$$

with H the local subgraph assigned by $\mathcal{H}_1$.

Corollary 3.1
$\tilde{c}_B(G)(v,H)$ is a local network function.

Proof:
This is a direct consequence of the fact that $\mathcal{H}_1$ is an environment generator (see Lemma 3.3). □

Lemma 3.5 Let $\mathcal{H}_1$ be an environment generator and $\tilde{c}_B$ the local network function as given by (3.15). Then $\tilde{c}_B$ approximates c_B with approximation error $\epsilon \leq S \cdot N_{\text{in}}(v) \cdot N_{\text{out}}(v)$.

Proof:
The approximation error ϵ of $\tilde{c}_B$ is given by the definitions in (3.10) and (3.15):

$$\begin{aligned}
\epsilon &= |c_B(G)(v) - \tilde{c}_B(G)(v,H)| \\
&= \left| \sum_{u,w \in V_G \setminus \{v\}} \frac{\sigma(u,v,w)}{\sigma(u,w)} - \sum_{\substack{v_0 \to v \to v_1 \\ v_0 \neq v_1}} \frac{\sigma(v_0,v,v_1)}{\sigma(v_0,v_1)} \right| \\
&= \sum_{u \not\to v \not\to w} \frac{\sigma(u,v,w)}{\sigma(u,w)} \\
&\approx \frac{1}{S} \sum_{\substack{u \xrightarrow{+} v_0 \to v \to v_1 \xrightarrow{+} w \\ v_0 \neq v, v_1 \neq v, u \neq w}} \sigma(u,v,w) \\
&\leq S \cdot N_{\text{in}}(v) \cdot N_{\text{out}}(v)
\end{aligned} \tag{3.16}$$

□

In the case of an undirected network, Eq. (3.16) rewrites to $\epsilon \leq S \cdot N(v)^2 = S \cdot n_{H_1}^2$.

Corollary 3.2
$\tilde{c}_B$ is not a local approximation.

Proof:
As the approximation error in Eq. 3.16 can be arbitrarily large depending on the specific degree of the vertex we consider as well as the expectation of the number of paths S, the error has no definite bound that depends on the size of the subgraph. As such, $\tilde{c}_B$ is not a local approximation. □

In contrast with closeness centrality, betweenness does not have a built-in mechanism that dampens contributions from vertices further away. Nonetheless, a decomposition shows that betweenness is proportional to an approximation consisting of vertices from H_1. Adding vertices further away will obviously lower the approximation error.

By extending the environment around v to cover neighbours more steps away (as with closeness centrality), the approximation becomes less dependent on the expectation value S. Hence, it is to be expected that $\mathcal{H}_d$, $d > 1$ will result in a more accurate order approximation. In Section 5 we show that such a local order approximation is indeed moderately successful.

3.3 PageRank centrality

The degree centrality of a vertex simply assigns to each vertex the number of neighbours it has as score. As such it is a fairly crude measure of centrality. By taking into account the centrality of the adjacent neighbours, we obtain a recursive improvement over degree centrality which is known as PageRank.

The *PageRank centrality* c_P of a vertex v is given by:

$$c_P(G)(v) = (1-\alpha) + \alpha \sum_{w \in V : w \to v} \frac{c_P(G)(w)}{k(w)} \quad . \tag{3.17}$$

The constant α is the so-called damping factor of the algorithm which corresponds to the probability of a random jump to another vertex. As the definition shows, it is a recursive measure that weighs vertices by importance of its in-neighbours, scaled by their outdegree.

If we are to solve Eq. (3.17) iteratively, we first rewrite it in matrix notation as

$$\mathbf{x}_{i+1} = \alpha \mathbf{M} \mathbf{x}_i + \frac{1-\alpha}{N} \mathbf{1} \ , \tag{3.18}$$

with $\mathbf{x}_i$ the PageRank vector at the i-th iteration, and $\mathbf{M}$ the stochastic link matrix derived from $\mathbf{A}$ as $\mathbf{M} = (\mathbf{K}^{-1}\mathbf{A})^T$. Here, $\mathbf{K}$ is the diagonal degree matrix and $\mathbf{A}$ once again the $N \times N$ adjacency matrix that corresponds to E_G. The PageRank vector is usually initialized as $\mathbf{x}_0 = \frac{1-\alpha}{N}\mathbf{1}$. Consequently, we have

$$\mathbf{x}_i = \sum_{j=0}^{i} (\alpha \mathbf{M})^j \left(\tfrac{1-\alpha}{N}\right) \mathbf{1} \tag{3.19}$$

$$= \left(\tfrac{1-\alpha}{N}\right) \left(\mathbf{I} - (\alpha \mathbf{M})^{i+1}\right) \cdot (\mathbf{I} - \alpha \mathbf{M})^{-1} \quad . \tag{3.20}$$

An intuitive approximation of PageRank $\tilde{c}_P$ consists of a finite number of such iterations. If we put this in the perspective of a single vertex, the approximation consists of taking vertices one step further away into account at each iteration.

However, the subgraph thus obtained is not a local subgraph. The reason for this is that PageRank is defined recursively on the *incoming* edges of neighbours of v. Such edges are in general not known from a local perspective. We will first examine the local approximation with the assumption that we have access to an index containing the incoming edges for each vertex. Thereafter we continue with a local approximation that uses a true local subgraph.

3.3.1 PageRank approximation with incoming edges

Note that if we consider only a single vertex, each subsequent multiplication in Eq. (3.20) corresponds to the set of (in-)neighbours one step further away. Let $\mathcal{I}_d$ be the function that assigns to each vertex the subgraph I consisting of vertices at most d in-edges away:

$$I(v) = \left\{ w \in V_G \mid w \xrightarrow{d} v \right\} \ , \tag{3.21}$$

then by restricting c_P to this environment we obtain an approximation of PageRank.

Lemma 3.6 $I(v)$ is not a local subgraph.

Proof:
$I(v)$ does not satisfy the third axiom as per Definition 2.1: given $w \in I(v)$ we know that $w \xrightarrow{*} v$, but in a directed graph this does not imply $v \xrightarrow{*} w$. □

Corollary 3.3

$\mathcal{I}_d$ is not an environment generator.

Proof:
This follows directly from Lemma 3.6. □

Nonetheless, we can use I_d as a restricted subgraph to approximate PageRank with, e.g. to calculate $\tilde{c}_P(G)(v, I_d)$.

Lemma 3.7 $\tilde{c}_P(G)(v, I_d)$ has approximation error $\epsilon \leq \alpha^{d+1} \cdot ||\mathbf{M}||^{d+1} \cdot ||\mathbf{x}_\infty||$.

Proof:
The approximation error is obtained by

$$\begin{aligned}\mathbf{x}_\infty - \mathbf{x}_d &= \sum_{j=0}^{\infty} (\alpha\mathbf{M})^j \left(\tfrac{1-\alpha}{N}\right)\mathbf{1} - \sum_{j=0}^{d} (\alpha\mathbf{M})^j \left(\tfrac{1-\alpha}{N}\right)\mathbf{1} \\ &= \left(\tfrac{1-\alpha}{N}\right) \sum_{j=d+1}^{\infty} (\alpha\mathbf{M})^j \\ &= (\alpha\mathbf{M})^{d+1}\mathbf{x}_\infty \ . \end{aligned} \tag{3.22}$$

The approximation error ϵ as a real number is obtained by taking the (Euclidean) norm of this difference:

$$\epsilon = ||\mathbf{x}_\infty - \mathbf{x}_d|| \leq \alpha^{d+1} \cdot ||\mathbf{M}||^{d+1} \cdot ||\mathbf{x}_\infty|| \ . \tag{3.23}$$

□

Corollary 3.4
$\tilde{c}_P(G)(v, I_d)$ is not a local approximation.

Proof:
This follows directly from the fact that $\mathcal{I}_d$ is not an environment generator (see Corollary 3.3). □

Note that if we choose $d = 1$, the approximation score is directly proportional to the in-degree $N_{\text{in}}(v)$. In Section 5 we demonstrate how this estimate is a good approximation of the true PageRank. However, as shown above, the subgraph assigned by $\mathcal{I}_d$ is not a local subgraph.

3.3.2 True local PageRank approximation

To work with a true local subgraph, we use the environment generator $\mathcal{H}_d$. Approximating PageRank is again accomplished by applying the recursion d times. However, the lack of knowledge of incoming edges has consequences for the approximation error. This error is strictly greater than with the use of $\mathcal{I}_d$ as an environment generator, as there are likely vertices that contribute towards the PageRank score that are not in the local subgraph. Vertices w in the subgraph have an edge towards v with a probability which can be expressed in terms of the reciprocity of the graph. In other words, if w is k steps away from v, then it has a probability ρ_k to have an edge to v directly (see Def. 2.2). This allows us to rewrite Eq. (3.17) as a local approximation:

$$\tilde{c}_P(G)(v, H_d) = (1-\alpha) + \alpha \sum_{k=1}^{d} \rho_k \left[\sum_{w \in H : v \xrightarrow{k} w} \frac{\tilde{c}_P(G)(w, H_d)}{k(w)} \right] \ . \tag{3.24}$$

In matrix form, we encapsulate the probability of reciprocal edges in the new definition of the matrix $\mathbf{M}$. Let $\hat{\mathbf{M}}_d = \sum_k^d \rho_k (\mathbf{K}^{-1}\mathbf{A}^k)^T$. Note that $\hat{\mathbf{M}}_d$ is not a stochastic matrix, as its rows do not (necessarily) sum to 1 anymore. Now the approximated vector after i iterations is given by

$$\tilde{c}_P(G)(v, H_d) = \tilde{\mathbf{x}}_d = \sum_{j=0}^{d} \left(\alpha\hat{\mathbf{M}}_d\right)^j \left(\tfrac{1-\alpha}{N}\right)\mathbf{1} \ . \tag{3.25}$$

Lemma 3.8 $\tilde{c}_P(G)(v, H_d)$ has approximation error

$$\epsilon \leq \alpha^{d+1} \cdot ||\mathbf{M}^{d+1}|| \cdot ||\mathbf{x}_\infty|| + \left(\tfrac{1-\alpha}{N}\right) \left[\frac{N - \alpha^{d+1}||\mathbf{M}||^{d+1}}{N - \alpha||\mathbf{M}||} - \frac{N - \alpha^{d+1}\rho_1^{d+1}||\mathbf{M}||^{d+1}}{N - \rho_1\alpha||\mathbf{M}||} \right] .$$

Proof:
The approximation error is obtained by

$$\mathbf{x}_\infty - \tilde{\mathbf{x}}_d = \sum_{j=0}^{\infty} (\alpha\mathbf{M})^j \left(\tfrac{1-\alpha}{N}\right) \mathbf{1} - \sum_{j=0}^{d} \left(\alpha\hat{\mathbf{M}}_d\right)^j \left(\tfrac{1-\alpha}{N}\right) \mathbf{1} \ . \tag{3.26}$$

If we assume that the probability of a reciprocal edge given a path of length 2 or more is negligible (which is the case for the directed networks we consider, see Table 3), then this is further reformulated as

$$\begin{aligned}
\mathbf{x}_\infty - \tilde{\mathbf{x}}_d &= \sum_{j=d+1}^{\infty} (\alpha\mathbf{M})^j \left(\tfrac{1-\alpha}{N}\right) \mathbf{1} + \sum_{j=0}^{d} (1 - \rho_1^j)(\alpha\mathbf{M})^j \left(\tfrac{1-\alpha}{N}\right) \mathbf{1} \\
&= \sum_{j=d+1}^{\infty} (\alpha\mathbf{M})^j \left(\tfrac{1-\alpha}{N}\right) \mathbf{1} + \sum_{j=0}^{d} (\alpha\mathbf{M})^j \left(\tfrac{1-\alpha}{N}\right) \mathbf{1} - \sum_{j=0}^{d} (\rho_1\alpha\mathbf{M})^j \left(\tfrac{1-\alpha}{N}\right) \mathbf{1} \\
&= (\alpha\mathbf{M})^{d+1}\mathbf{x}_\infty \\
&\quad + \left(\tfrac{1-\alpha}{N}\right) \left[(\mathbf{1} - (\alpha\mathbf{M})^{d+1}) \cdot (\mathbf{1} - \alpha\mathbf{M})^{-1} - (\mathbf{1} - (\rho_1\alpha\mathbf{M})^{d+1}) \cdot (\mathbf{1} - \rho_1\alpha\mathbf{M})^{-1} \right] .
\end{aligned} \tag{3.27}$$

Consequently, with the shorthand $y = \frac{\alpha||\mathbf{M}||}{N}$,

$$\begin{aligned}
\epsilon &= ||\mathbf{x}_\infty - \tilde{\mathbf{x}}_d|| \\
&\leq \alpha^{d+1} \cdot ||\mathbf{M}||^{d+1} \cdot ||\mathbf{x}_\infty|| + \tfrac{(1-\alpha)}{N} \left[\frac{N - \alpha^{d+1}||\mathbf{M}||^{d+1}}{N - \alpha||\mathbf{M}||} - \frac{N - \alpha^{d+1}\rho_1^{d+1}||\mathbf{M}||^{d+1}}{N - \rho_1\alpha||\mathbf{M}||} \right] \\
&= \alpha^{d+1} \cdot ||\mathbf{M}||^{d+1} \cdot ||\mathbf{x}_\infty|| + (1-\alpha) \frac{(1-\rho_1)y + (\rho_1^{d+1} - 1)(y)^{d+1} + (\rho_1 - \rho_1^{d+1})y^{d+2}}{(1-y)(1-\rho_1 y)} \ .
\end{aligned} \tag{3.28}$$

□

Note that the first term in this equation is exactly the approximation error in Eq. (3.23). This error can be seen as the *convergence error*, since this part of the deviation stems from the fact that only a fixed number of iterations are executed. The second term can be seen as the *model error*, because this part occurs as consequence of the lack of knowledge about incoming edges. It can be arbitrarily large for small values of ρ_1. However, when ρ_1 approaches 1, Eqs. (3.28) and (3.23) become equivalent. In this scenario, only the convergence error remains.

Corollary 3.5
Let $\mathcal{H}_d$ be an environment generator, then $\tilde{c}_P(G)(v, H_d)$ is not a local approximation of $c_P(G)(v)$.

Proof:
This follows directly from the fact that the approximation error of $\tilde{c}_P(G)(v, H_d)$ does not have a clear bound depending on the size of the subgraph H_d. □

In Section 5 we show that $\tilde{c}_P(G)(v, I_d)$ successfully approximates the PageRank order, while $\tilde{c}_P(G)(v, H_d)$ is unfortunately not able to.

4 Complexity

The local approximation approach tacitly assumes that the centrality scores for only a relatively small number of vertices is required for a given application. In such a scenario, calculating the entire vector of centrality scores would be a waste of resources. However, because of the low time complexity of the local approximation algorithms, local approximations might even be faster in the case where entire vector must be obtained.

4.1 Closeness

The calculation of the true closeness centrality for all vertices in the network requires knowledge of all shortest paths. This can be obtained through several algorithms, such as the Floyd-Warshall algorithm with running time complexity $O(n^3)$ or Dijkstra's algorithm which has time complexity $O(nm + n^2 \log n)$, depending on the implementation of its priority queue.

The local approximation of closeness centrality considers for each vertex the number of neighbours at most d steps away and therefore has running time complexity $O(nk^d)$, with k the average degree. Note that $nk = m$, so our approximation based on $\mathcal{H}_2$ is calculated in $O(mk)$ time. As most real world graphs are sparse, $k \ll n$, so the local approximation significantly improves calculation time.

4.2 Betweenness

Efficient algorithms for the calculation of betweenness centrality have been proposed by Brandes [9] and Newman [10], both with time complexity $O(nm)$. To analyze the complexity of the approximation, we take a closer look at its definition in Eq. (3.14), which we repeat here for the sake of convenience:

$$\sum_{\substack{v_0 \to v \to v_1 \\ v_0 \neq v_1}} \frac{\sigma(v_0, v, v_1)}{\sigma(v_0, v_1)} \ .$$

This definition coincides with the concept of *ego centrality*, which was discussed by Everett and Borgatti [11]. In the same paper, Everett and Borgatti show that ego centrality can quickly be calculated by summing the reciprocal of the entries of $\mathbf{E}^2$, where $\mathbf{E}$ is the adjacency matrix containing v and its neighbors, hence the local subgraph contains $k + 1$ vertices. Using the Coppersmith-Winograd algorithm for matrix multiplication, the ego centrality is calculated in $O((k+1)^{2.376})$ time. Once again, k denotes the average degree of the network. Computing the ego centrality for all vertices thus leads to a running time complexity of $O(n(k+1)^{2.376})$. In general, $(k+1)^{2.376} \ll m$, which implies that the local approximation is faster than the actual betweenness centrality, even when all vertices are considered.

However, when vertices at further distance from v are considered (i.e. for $\mathcal{H}_d$ with $d > 1$) or the network is directed, the fast calculation as suggested by Everett and Borgatti can no longer be used. In this situation, one of the algorithms from Brandes or Newman should be used on the local subgraph instead. An upper bound on the number of vertices in the local subgraph is $\sum_{i=0}^{d} k^i$, which is of the order of magnitude k^d. In the worst-case scenario all these vertices are connected in a tree structure, which gives at most $k^d - 1$ edges. Applying one of the betweenness centrality algorithms is therefore calculated in $O(k^{2d})$, which must be repeated for all vertices in the network, leading to a total running time complexity of $O(nk^{2d})$. In this case we have a trade-off between the size of the local subgraph and the computation time; the local approximation is more efficient iff $k^{2d} < m$.

4.3 PageRank

The problem of finding the PageRank vector corresponds to finding the principal eigenvector of the matrix $\mathbf{M} = (\mathbf{K}^{-1}\mathbf{A})^T$. In practice, this is done using a straightforward iterative algorithm, which is applied until a certain convergence threshold is reached. Let γ be the number of iterations

Table 1: Small sized networks. Given are the number of vertices n, the number of edges m and the diameter of the graph D.

Network	Label	Edges	n	m	D
Bottlenose dolphins [12]	BD	undirected	62	159	5
Football players [13]	FP	undirected	115	613	8
Political books recommendations [14]	PBR	undirected	105	441	8
Les Miserables [15]	LM	undirected	77	254	5

needed until convergence is satisfied, then the running time complexity of iterative PageRank is $O(\gamma nk) = O(\gamma m)$ – which is by far the quickest to calculate centrality measure we consider, given reasonable values for γ. Calculating the local approximation of PageRank simply iterates d times, using only vertices in $\mathcal{H}_d$. The number of vertices in this subgraph is at most k^d as we have shown before, so the complexity of the approximation of a single vertex' score is $O(dk^{d+1})$. Obtaining the entire approximated PageRank vector is then done in $O(dnk^{d+1})$. In contrast to closeness and betweenness, the approximation of the entire PageRank vector is actually slower than applying iterative PageRank (once again assuming reasonable values for γ and d). The rationale behind this is the fact that in the approximation many of the local subgraphs will overlap, but each is considered separately in the local approximation.

Nonetheless, when instead of the entire PageRank vector only the scores for a subset of the vertices is required, the local approximation is definitely a faster alternative.

5 Experimental Setup

In Section 3 we analyzed the approximation errors of local approximations. In the case of local order approximations however, we only consider the ordering that the measures induce. To investigate how the suggested local approximations perform as local order approximations, we calculated the exact values of the centrality measures and compared these to those obtained by a local approximation.

5.1 Small networks

As our initial experiment, we considered four small, undirected, networks from Mark Newman's online collection[1]. Details of these networks can be found in Table 1.

Since these networks are fairly small, we can plot the values for the true measure and its approximation in a chart. This is done in Figures 1, 2, 3, 4 and 5. For each of these charts, the true centrality score is shown on the horizontal axis. On the vertical axis is the local approximation of the centrality measure. If the local approximation is successful, all points should lie on the diagonal, as all values for the global and the local measure would be the same. In addition, the axes would have the same scale. In the case of a local order approximation, the points should lie on the diagonal as well, but the axes may differ.

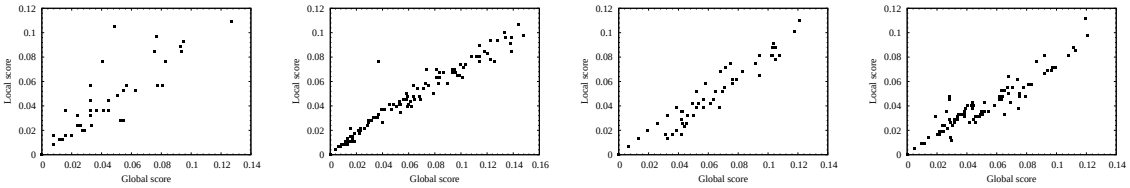
5.2 Larger networks

In addition to these small examples, we experimented on several larger graphs taken from the Stanford Large Network Collection[2]. The collection contains both directed and undirected networks. Details of the selected networks can be found in Table 2.

When considering significantly larger networks, we cannot simply plot the measures on a chart for obvious reasons. We therefore consider only the rank correlation coefficients between the rankings induced by the centrality measure and its approximation. Tables 6a, 6b and 6c show

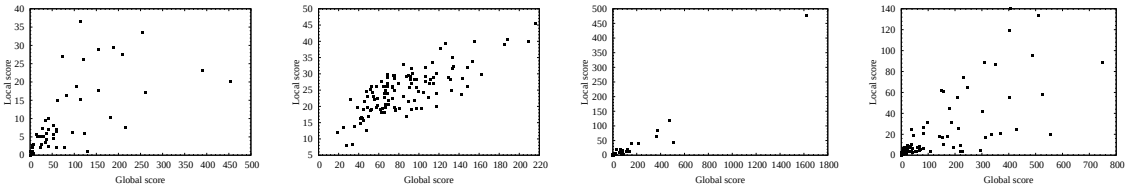
[1] `http://www-personal.umich.edu/~mejn/netdata/`

[2] `http://snap.stanford.edu/data/`



(a) BD (b) FP (c) LM (d) PBR

Figure 1: Closeness centrality. All local subgraphs are generated by $\mathcal{H}_2$.

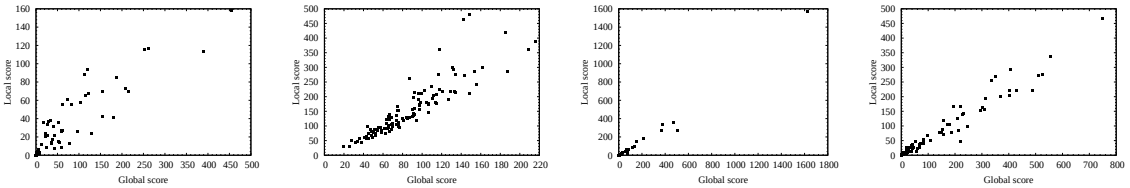


(a) BD (b) FP (c) LM (d) PBR

Figure 2: Betweenness centrality. All local subgraphs are generated by $\mathcal{H}_1$.

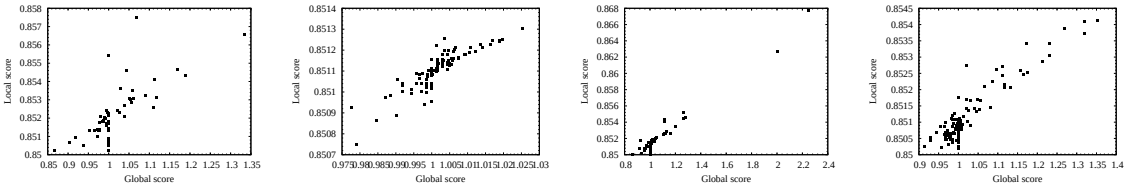
Kendall τ for several configurations. However, there is an important caveat when considering correlation coefficients for larger networks. As can be observed from the charts for the smaller networks, the centrality measures (and their approximations) yield many scores close to or exactly zero. This in itself is not surprising nor incorrect; dangling vertices will have zero shortest paths passing through them and hence c_B will be zero. However, when calculating correlation, these vertices will have the same rank in both the approximation as well as the actual measure. Again, this is as intended, but may lead to premature conclusions about the quality of the local order approximation. For example, a local approximation $\forall_v \hat{c}(v) = 0$ may show strong correlation, simply because many zero values exist in the true result.

To avoid this pitfall, we also display the correlation for each configuration considering only vertices for which at least one component (the global or the local measure) is nonzero. These scores are shown in Figures 6a and 6b as $\tau\prime$. In addition, the table lists the percentage of the vertices that remain when vertices with zero scores have been ignored, indicated by $\%\prime$. Note that this problem does not apply to PageRank, as it does not yield zero values.



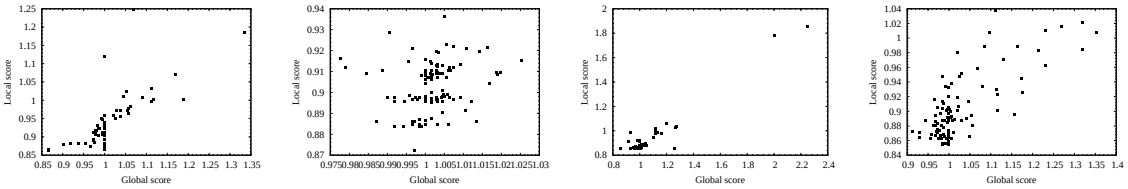
(a) BD (b) FP (c) LM (d) PBR

Figure 3: Betweenness centrality. All local subgraphs are generated by $\mathcal{H}_2$.



(a) BD (b) FP (c) LM (d) PBR

Figure 4: PAGERANK CENTRALITY. All local subgraphs are generated by $\mathcal{H}_1$.



(a) BD (b) FP (c) LM (d) PBR

Figure 5: PAGERANK CENTRALITY. All local subgraphs are generated by $\mathcal{H}_2$.

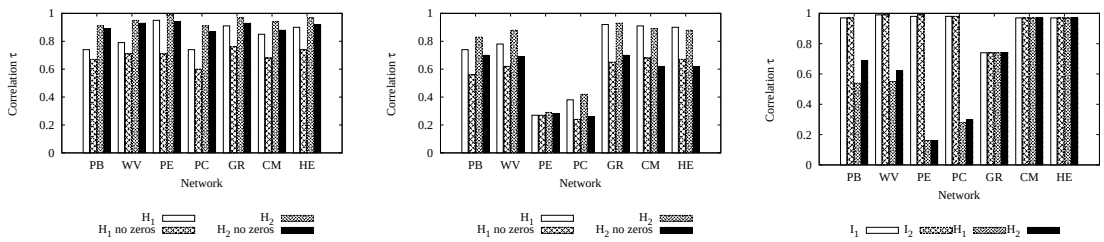
5.3 Results

5.3.1 Closeness

The results as shown in the series of plots as well as the correlation values reveal a number of facets of local approximations. First, we note that closeness centrality is very well suited for local approximation. We attribute this fact to the damping factor in its definition. Since vertices further away contribute increasingly less to the closeness score, ignoring vertices further than d steps away does not have an all too great impact on the result. Not only can the order induced by c_C be approximated locally, the approximation strategy also applies to the actual values assigned. This allows us to conclude that the approximation of closeness as given by Eq. (3.7) is both a local approximation as well as a local order approximation.

Table 2: Large sized networks. Given are the number of vertices n, the number of edges m and the diameter of the graph D. In addition the shorthand that is used throughout this paper and the directionality of the edges are listed.

Network	Label	Edges	n	m	D
Political blogs [16]	PB	directed	1210	18139	9
Wikipedia user votes [17]	WV	directed	7115	103689	7
Pages linking to www.epa.gov [18]	PE	directed	4271	8965	9
Pages matching query "california" [19]	PC	directed	6175	16150	15
Arxiv General Relativity Collaboration [20]	GR	undirected	5241	28968	17
Arxiv Condensed Matter Collaboration [20]	CM	undirected	23133	186878	15
Arxiv High Energy Physics Collaboration [20]	HE	undirected	9877	51971	17



(a) Closeness. (b) Betweenness. (c) PageRank.

Figure 6: Correlation between true order and local order approximation for different local subgraphs. For closeness and betweenness the correlation for non-zero scores is shown; for PageRank the correlation for $\mathcal{I}_d$ is shown.

Table 3: Reciprocity-at-distance-k $\rho_k(G)$ for directed networks.

Network	$\rho_1(G)$	$\rho_2(G)$
PB	0.22	0.03
WV	0.06	0.01
PE	0.01	0.00
PC	0.02	0.00

5.3.2 Betweenness

Second, we observe that for undirected networks the order approximation of betweenness correlates strongly with the actual scores. Since local algorithms cannot use incoming edges, we expected that approximating betweenness on directed graphs would perform significantly less. This is indeed the case for most of the larger networks we consider. However, for some networks the order induced by betweenness is actually fairly well approximated by our suggestion in Eq. (3.15). This is a surprising result, and as such we took a closer look at the differences between the networks on which the approximation scores highly and those where it does not. We expected that the success of betweenness approximation on directed graphs depends on the reciprocity-at-distance-k $\rho_k(G)$. Table 3 shows the scores for $\rho_1(G)$ and $\rho_2(G)$ for the directed graphs we considered. Indeed, the PB network and the WV network (for which the betweenness approximation is fairly successful) have significantly larger reciprocity than the other networks.

5.3.3 PageRank

Third, the results for the approximation of PageRank show that when we assume to have an index available to request incoming edges, considering only the incoming neighbours of one or two steps away correlates very strongly with the actual PageRank. This applies to both directed and (obviously) undirected networks. In the case of the true local approach where such an index is unavailable, the correlation for directed graphs drops significantly. Here we observe a pattern similar to betweenness, where the correlation with the approximation increases with the increased fraction of reciprocal edges.

6 Conclusion

In this paper we have presented a framework of definitions that can be used to find local approximations of network functions. These local approximations consider only a subgraph of vertices around a specific vertex. The benefit of this local approach is that on large networks, calculation

on small subgraphs is significantly faster than obtaining the exact scores. However, the technique intentionally throws away much information of the network connectivity, and consequently introduces an approximation error. For three examples we considered expressions for these errors. In addition, we analyzed the computational benefits of the local approximations, and their performance on real networks.

From the analysis and the experiments we conclude that some measures are well suited for local approximations. The best example is closeness centrality, which has a local approximation, a local order approximation, an approximation error which is clearly defined in terms of the parameters of the local subgraph and which performs very well on real networks. Since closeness centrality is used in disciplines that consider large networks, such as neuroscience [21], the local approximation has great application perspective. The same applies to betweenness centrality, although here it should be noted that the approximation is much better on undirected graphs. However, for directed graphs with some degree of reciprocity in the edges, the local order approximation of betweenness centrality performs remarkably well. Finally, PageRank can be approximated very well on undirected networks, or on directed networks if we assume that incoming edges are known. Although this is not the case in general, large scale search engines may benefit from local order approximations, since they usually have indices of incoming edges available.

References

[1] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[2] David Bader, Shiva Kintali, Kamesh Madduri, and Milena Mihail. Approximating betweenness centrality. pages 124–137. 2007.

[3] Ulrik Brandes and Christian Pich. Centrality estimation in large networks. *I. J. Bifurcation and Chaos*, 17(7):2303–2318, 2007.

[4] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, New York, NY, USA, 2006. ACM.

[5] M. Hinne. Local identification of web graph communities. In *ICTIR '07: Proceedings of the 1st International Conference on the Theory of Information Retrieval*, pages 261–278, Budapest, Hungary, October 2007. Alma Mater Series.

[6] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[8] Stefano Allesina and Mercedes Pascual. Googling food webs: Can an eigenvector measure species' importance for coextinctions? *PLoS Comput Biol*, 5(9):e1000494, 09 2009.

[9] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[10] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(1):016132, Jun 2001.

[11] Everett and Borgatti. Ego network betweenness. *Social Networks*, 27(1):31–38, January 2005.

[12] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, and S.M. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. can geographic isolation explain this unique trait? *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

[13] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, June 2002.

[14] V. Krebs. Political polarization during the 2008 us presidential campaign.

[15] D. E. Knuth. The stanford graphbase: A platform for combinatorial computing. Addison-Wesley, Reading, MA, 1993.

[16] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pages 36–43, New York, NY, USA, 2005. ACM.

[17] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1361–1370, New York, NY, USA, 2010. ACM.

[18] Jon Kleinberg. Pages linking to www.epa.gov. http://www.cs.cornell.edu/courses/cs685/2002fa/.

[19] Jon Kleinberg. Pages matching the query "california". http://www.cs.cornell.edu/courses/cs685/2002fa/.

[20] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1, March 2007.

[21] Olaf Sporns. *Networks of the Brain*. The MIT Press, 2010.