HAL
archives-ouvertes.fr

# Metropolis-Hastings Algorithms for Estimating Betweenness Centrality Talel Abdessalem

Mostafa Haghir Chehreghani, Talel Abdessalem, Albert Bifet

▶ **To cite this version:**

Mostafa Haghir Chehreghani, Talel Abdessalem, Albert Bifet. Metropolis-Hastings Algorithms for Estimating Betweenness Centrality Talel Abdessalem. 22nd International Conference on Extending Database Technology EDBT 2019, Mar 2019, Lisbon, Portugal. 10.5441/002/edbt.2019.87 . hal-02339557

**HAL Id: hal-02339557**
**https://hal.telecom-paris.fr/hal-02339557**

Submitted on 30 Oct 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Short Paper

# Metropolis-Hastings Algorithms for Estimating Betweenness Centrality

Mostafa Haghir Chehreghani
LTCI, Télécom ParisTech
Paris
mostafa.chehreghani@gmail.com

Talel Abdessalem
LTCI, Télécom ParisTech
Paris
talel.abdessalem@
telecom-paristech.fr

Albert Bifet
LTCI, Télécom ParisTech
Paris
albert.bifet@telecom-paristech.fr

## ABSTRACT

Recently, an optimal probability distribution was proposed to sample vertices for estimating betweenness centrality, that yields the minimum approximation error. However, it is computationally expensive to directly use it. In this paper, we investigate exploiting Metropolis-Hastings technique to sample based on this distribution. As a result, first given a network $G$ and a vertex $r \in V(G)$, we propose a Metropolis-Hastings MCMC algorithm that samples from the space $V(G)$ and estimates betweenness score of $r$. The stationary distribution of our MCMC sampler is the optimal distribution. We also show that our MCMC sampler provides an $(\epsilon, \delta)$-approximation. Then, given a network $G$ and a set $R \subset V(G)$, we present a Metropolis-Hastings MCMC sampler that samples from the joint space $R$ and $V(G)$ and estimates relative betweenness scores of the vertices in $R$. We show that for any pair $r_i, r_j \in R$, the ratio of the expected values of the estimated relative betweenness scores of $r_i$ and $r_j$ with respect to each other is equal to the ratio of their betweenness scores. We also show that our joint-space MCMC sampler provides an $(\epsilon, \delta)$-approximation of the relative betweenness score of $r_i$ with respect to $r_j$.

## 1 INTRODUCTION

*Centrality* is a structural property of vertices (or edges) in a network that quantifies their relative importance. For example, it determines the importance of a person within a social network, or a road within a road network. Freeman [14] introduced and defined *betweenness centrality* of a vertex as the number of shortest paths from all (source) vertices to all others that pass through that vertex. He used it for measuring the control of a human over the communications among others in a social network [14]. Betweenness centrality is also used in some well-known algorithms for clustering and community detection in social and information networks [16].

Although there exist polynomial time and space algorithms for betweenness centrality computation, the algorithms are expensive in practice. Currently, the most efficient existing exact method is Brandes's algorithm [5]. Time complexity of this algorithm is $O(nm)$ for unweighted graphs and $O(nm + n^2 \log n)$ for weighted graphs with positive weights, where $n$ and $m$ are the number of vertices and the number of edges of the network, respectively. This means exact betweenness centrality computation is not applicable, even for mid-size networks. However, there exist observations that may improve the computation of betweenness scores in practice.

© 2019 Copyright held by the owner/author(s). Published in Proceedings of the 22nd International Conference on Extending Database Technology (EDBT), March 26-29, 2019, ISBN 978-3-89318-081-3 on OpenProceedings.org.
Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.

- First, in several applications it is sufficient to compute betweenness score of only one or a few vertices. For instance, this index might be computed for only core vertices of communities [23] in social/information networks or for only hubs in communication networks. Chehreghani [9] has discussed some situations where it is required to compute betweenness score of only one vertex. Note that these vertices are not necessarily those that have the highest betweenness scores. Hence, algorithms that identify vertices with the highest betweenness scores [21] are not applicable. While exact computation of this index for one vertex is not easier than that for all vertices, Chehreghani [9] and later Riondato and Kornaropoulos [21] respectively showed that this index can be estimated more effectively for one arbitrary vertex and for $k$ vertices that have the highest scores.
- Second, in practice, instead of computing betweenness scores, it is usually sufficient to *compute betweenness ratios* or *rank vertices* according to their betweenness scores [21]. For example, Daly and Haahr [12] exploited betweenness ratios for finding routes that provide good delivery performance and low delay in Mobile Ad hoc Networks. The other application is handling cascading failures [1].

While the above mentioned observations do not yield a better algorithm when exact betweenness scores are used, they may improve approximate algorithms. In the current paper, we exploit both of these observations to design more effective approximate algorithms. In the first problem studied in this paper, we assume that we are given a vertex $r \in V(G)$ and we want to estimate its betweenness score. In the second problem, we assume that we are given a set $R \subset V(G)$ and we want to estimate the ratios of betweenness scores of vertices in $R$. The second problem is formally defined as follows: given a graph $G$ and a set $R \subset V(G)$, for any two vertices $r_i$ and $r_j$ in $R$, we want to estimate the *relative betweenness score* of $r_i$ with respect to $r_j$, denoted by $BC_{r_j}(r_i)$ (see Equation 8 of Section 4.3 for the formal definition of relative betweenness score). The ratio of the expected values of our estimations of $BC_{r_j}(r_i)$ and $BC_{r_i}(r_j)$ is equal to the ratio of betweenness scores of $r_i$ and $r_j$.

In [9], Chehreghani presented the optimal probability distribution for estimating betweenness centrality, that yields the minimum approximation error 0. However, this distribution cannot be directly used, as computing the constant factor of probability densities is computationally expensive. A natural solution for this problem is to use Metropolis-Hastings sampling [18]. In this short paper, we investigate the possibility of using such a sampling method for the two aforementioned problems and theoretically analyze the resulted algorithms. More precisely, our key contributions are as follows.

- Given a graph $G$ and a vertex $r \in V(G)$, in order to estimate betweenness score of $r$, we develop an MCMC sampler

that samples from the space $V(G)$. Unlike existing work, our samples are non-iid and the stationary distribution of our MCMC sampler is the *optimal probability distribution* [9]. We also show that our MCMC sampler provides an $(\epsilon, \delta)$-approximation of the betweenness score of $r$ ($\epsilon \in \mathbb{R}^+$ and $\delta \in (0, 1)$).

- Given a graph $G$ and a set $R \subset V(G)$, in order to estimate relative betweenness scores of all pairs of vertices in $R$, we develop an MCMC sampler that samples from the joint space $R$ and $V(G)$. This means each sample (state) in our MCMC sampler is a pair $\langle \mathfrak{r}, \mathfrak{v} \rangle$, where $\mathfrak{r} \in R$ and $\mathfrak{v} \in V(G)$. For any two vertices $r_i, r_j \in R$, we show that our joint-space MCMC sampler provides an $(\epsilon, \delta)$-approximation of the relative betweenness score of $r_i$ with respect to $r_j$.

Techniques similar to our algorithm (the second algorithm that estimate *relative* betweenness centrality) have already been used in *statistical physics* to estimate *free energy differences* [3]. However, they are new in the context of network analysis. Our current work takes the first step in bridging these two domains. This step can be further extended by proposing algorithms similar to our work for estimating other network indices. As a result, a novel family of techniques might be introduced to the field of network analysis. We leave efficient implementations of our proposed algorithms and evaluating their empirical efficiency for future work.

## 2 PRELIMINARIES

Throughout the paper, $G$ refers to a graph (network). For simplicity and without loss of generality, we assume that $G$ is an undirected, connected and loop-free graph without multi-edges. Also, we assume that $G$ is an unweighted graph, unless it is explicitly mentioned that $G$ is weighted. $V(G)$ and $E(G)$ refer to the set of vertices and the set of edges of $G$, respectively. For a vertex $v \in V(G)$, by $G \setminus v$ we refer to the set of connected graphs generated by removing $v$ from $G$. A *shortest path* between two vertices $u, v \in V(G)$ is a path whose size is minimum, among all paths between $u$ and $v$. For two vertices $u, v \in V(G)$, we use $d(u, v)$, to denote the size (the number of edges) of a shortest path connecting $u$ and $v$. By definition, $d(u, u) = 0$ and $d(u, v) = d(v, u)$. For $s, t \in V(G)$, $\sigma_{st}$ denotes the number of shortest paths between $s$ and $t$, and $\sigma_{st}(v)$ denotes the number of shortest paths between $s$ and $t$ that also pass through $v$. *Betweenness centrality* of a vertex $v$ is defined as:

$$BC(v) = \frac{1}{|V(G)| \cdot (|V(G)| - 1)} \sum_{s,t \in V(G) \setminus \{v\}} \frac{\sigma_{st}(v)}{\sigma_{st}}.$$

A notion which is widely used for counting the number of shortest paths in a graph is the directed acyclic graph (DAG) containing all shortest paths starting from a vertex $s$ (see e.g., [5]). In this paper, we refer to it as the *shortest-path-DAG*, or *SPD* in short, rooted at $s$. For every vertex $s$ in graph $G$, the *SPD* rooted at $s$ is unique, and it can be computed in $O(|E(G)|)$ time for unweighted graphs and in $O(|E(G)| + |V(G)| \log |V(G)|)$ time for weighted graphs with positive weights [5]. Brandes [5] introduced the notion of the *dependency score* of a vertex $s \in V(G)$ on a vertex $v \in V(G) \setminus \{s\}$, which is defined as: $\delta_{s\bullet}(v) = \sum_{t \in V(G) \setminus \{v,s\}} \delta_{st}(v)$, where $\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$. We have: $BC(v) = \frac{1}{|V(G)| \cdot (|V(G)|-1)} \sum_{s \in V(G) \setminus \{v\}} \delta_{s\bullet}(v)$.

A Markov chain is a sequence of dependent random variables (states) such that the probability distribution of each variable given the past variables depends only on the last variable. An MCMC has *stationary distribution* if the conditional distribution of the $k + 1^{\text{th}}$ state given the $k^{\text{th}}$ state does not depend on $k$. Let $\mathbb{P}[x]$ be a probability distribution defined on the random variable $x$. When the function $f(x)$, which is proportional to the density of $\mathbb{P}[x]$, can be efficiently computed, the Metropolis-Hastings algorithm is used to draw samples from $\mathbb{P}[x]$. In a simple form (with symmetric proposal distribution), the Metropolis-Hastings algorithm first chooses an arbitrary initial state $x_0$. Then, iteratively: i) let $x$ be the current state. It generates a candidate $x'$ using the *proposal distribution* $q(x'|x)$, and ii) it moves from $x$ to $x'$ with probability $\min\left\{1, \frac{f(x')}{f(x)}\right\}$. The *proposal distribution* $q(x'|x)$ defines the conditional probability of proposing a state $x'$ given the state $x$. In the *Independence Metropolis-Hastings algorithm*, $q(x'|x)$ is independent of $x$, i.e., $q(x'|x) = g(x')$.

## 3 RELATED WORK

Brandes [5] introduced an efficient algorithm for computing betweenness centrality of a vertex, which is performed in $O(|V(G)||E(G)|)$ and $O(|V(G)||E(G)| + |V(G)|^2 \log |V(G)|)$ times for unweighted and weighted networks with positive weights, respectively. Çatalyürek et. al. [7] presented the *compression* and *shattering* techniques to improve efficiency of Brandes's algorithm for large graphs. The two natural extension of betweenness centrality to sets of vertices are *group betweenness centrality* [13] and *co-betweenness centrality* [8]. Brandes and Pich [6] and Bader et.al. [2] proposed approximate algorithms based on selecting $k$ source vertices and computing dependency scores of them on the other vertices in the graph. To estimate betweenness score of vertex $v$, Chehreghani [9] presented a non-uniform sampler, defined as follows: $\mathbb{P}[s] = \frac{1/d(v,s)}{\sum_{u \in V(G) \setminus \{v\}} 1/d(v,u)}$, where $s \in V(G) \setminus \{v\}$. Similar to these algorithms, our proposed algorithms are *source vertex samplers*, too. However, they use a new mechanism for sampling which is based on the Metropolis-Hastings algorithm. Riondato and Upfal [22] introduced a *pair sampler* for estimating betweenness scores of all (or top-$k$) vertices in a graph. Riondato and Kornaropoulos [21] and Borassi and Natale [4] presented *shortest path samplers* for estimating betweenness centrality of all vertices or the $k$ vertices that have the highest betweenness scores. The algorithm of [4] uses balanced bidirectional BFS (bb-BFS) to sample shortest paths. In bb-BFS, a BFS is performed from each of the two endpoints $s$ and $t$, in such a way that they are likely to explore about the same number of edges. Finally, Chehreghani et.al. [11] presented exact and approximate algorithms for computing betweenness centrality in directed graphs.

## 4 MCMC ALGORITHMS FOR ESTIMATING BETWEENNESS CENTRALITY

In this section, we present our MCMC sampler for estimating betweenness score of a single vertex; and our joint-space MCMC sampler for estimating relative betweenness scores of vertices in a given set.

### 4.1 Betweenness centrality as a probability distribution

Chehreghani [9] presented a randomized algorithm that admits a probability mass function as an input parameter. Then, he proposed an optimal sampling technique that computes betweenness score of a vertex $r \in V(G)$ with error 0. In optimal sampling, each

vertex $v$ is chosen with probability

$$\mathbb{P}_r[v] = \frac{\delta_{v\bullet}(r)}{\sum_{v' \in V(G)} \delta_{v'\bullet}(r)} \quad (1)$$

In other words, for estimating betweenness score of vertex $r$, each source vertex $v \in V(G)$ whose dependency score on $r$ is greater than 0, is chosen with probability $\mathbb{P}[v]$ defined in Equation 1.

In the current paper, for $r \in V(G)$ we want to estimate $BC(r)$ and also for all pairs of vertices $r_i, r_j$ in a set $R \subset V(G)$, the ratios $\frac{BC(r_i)}{BC(r_j)}$. For this purpose, we follow a *source vertex sampling* procedure where for each vertex $r$, we consider $\mathbb{P}_r[\cdot]$ defined in Equation 1 as the target probability distribution used to sample vertices $v \in V(G)$. It is, however, computationally expensive to calculate the normalization constant $\sum_{v' \in V(G)} \delta_{v'\bullet}(r)$ in Equation 1, as it gives the betweenness score of $r$. However, for two vertices $v_1, v_2 \in V(G)$, it might be feasible to compute the ratio $\frac{\mathbb{P}_r[v_1]}{\mathbb{P}_r[v_2]} = \frac{\delta_{v_1\bullet}(r)}{\delta_{v_2\bullet}(r)}$, as it can be done in $O(|E(G)|)$ time for unweighted graphs and in $O(|E(G)| + |V(G)| \log |V(G)|)$ time for weighted graphs with positive weights. This motivates us to propose Metropolis-Hastings sampling algorithms that for a vertex $r$, sample each vertex $v \in V(G)$ with the probability distribution $\mathbb{P}_r[v]$ defined in Equation 1.

## 4.2 A single-space MCMC sampler

In this section, we propose an MCMC sampler, defined on the space $V(G)$, to estimate betweenness centrality of a single vertex $r$. Our MCMC sampler consists of the following steps:

- First, we choose a vertex $\mathfrak{v}_0 \in V(G)$, as the initial state, uniformly at random.
- Then, at each iteration $t$, $1 \le t \le T$:
  - Let $\mathfrak{v}(t)$ be the current state of the chain.
  - We choose vertex $\mathfrak{v}'(t) \in V(G)$, uniformly at random.
  - With probability $\min\left\{1, \frac{\delta_{\mathfrak{v}'(t)\bullet}(r)}{\delta_{\mathfrak{v}(t)\bullet}(r)}\right\}$ we move from state $\mathfrak{v}(t)$ to the state $\mathfrak{v}'(t)$.

The sampler is an iterative procedure where at each iteration $t$, one transition may occur in the Markov chain. Let $M$ be the multi-set (i.e., the set where repeated members are allowed) of samples (states) accepted by our sampler. In the end of sampling, betweenness score of $r$ is estimated as

$$\ddot{BC}(r) = \frac{1}{(T+1)(|V(G)|-1)} \sum_{v \in M} \sum_{u \in V(G) \setminus \{v\}} \frac{\sigma_{vu}(r)}{\sigma_{vu}}. \quad (2)$$

This estimation does not give an unbiased estimation of $BC(r)$, however as we discuss below, by increasing $T$, $\ddot{BC}(r)$ can become arbitrarily close to $BC(r)$. In the rest of this section, we show that our MCMC sampler provides an $(\epsilon, \delta)$-approximation of $BC(r)$, where $\epsilon \in \mathbb{R}^+$ and $\delta \in (0, 1)$.

Theorem 4.1. *Let $\overline{\delta(r)}$ be the average of dependency scores of vertices in $V(G)$ on $r$, i.e., $\overline{\delta(r)} = \frac{\sum_{v \in V(G)} \delta_{v\bullet}(r)}{|(V(G)|}$, and $\Delta(r)$ be the maximum dependency score that a vertex in $G$ has on $r$. Let also $\mu(r)$ denote $\frac{\Delta(r)}{\overline{\delta(r)}}$. Then, for a given $\epsilon \in \mathbb{R}^+$, by our MCMC sampler and starting from any arbitrary initial state, we have*

$$\mathbb{P}\left[|\ddot{BC}(r) - BC(r)| > \epsilon\right] \le 2 \exp\left\{-\frac{T}{2}\left(\frac{2\epsilon(|V(G)|-1)}{\mu(r)\Delta(r)} - \frac{3}{T}\right)^2\right\}. \quad (3)$$

Due to space limitations, in this short paper we omit all the proofs. However, the interested reader may find them in a longer version of this text in [10]. In general, our $(\epsilon, \delta)$-approximation proofs are based on a theorem presented in [17] for the concentration analysis of MCMC samples and a theorem presented in [19] for the uniformly ergodicity of Independence Metropolis-Hastings algorithms.

Note that Inequality 3 does not depend on the initial state. Furthermore, in Inequality 3 it is not required to discard an initial part of the chain, called *burn-in*. More details on this can be found in [10]. $T$ is usually large enough so that we can approximate $\frac{3}{T}$ by 0. Hence, Inequality 3 yields that for given values $\epsilon \in \mathbb{R}^+$ and $\delta \in (0, 1)$, if $T$ is chosen such that

$$T \ge \frac{\mu(r)^2 \Delta(r)^2}{2\epsilon^2(|V(G)|-1)^2} \ln \frac{2}{\delta} \quad (4)$$

our MCMC sampler will estimate the betweenness score of $r$ within an additive error $\epsilon$ with a probability at least $1 - \delta$.

## 4.3 A joint-space MCMC sampler

In this section, we present an MCMC sampler to estimate the ratios of betweenness scores of the vertices in a set $R \subset V(G)$. Each state of this sampler is a pair $(\mathfrak{r}, \mathfrak{v})$, where $\mathfrak{r} \in R$ and $\mathfrak{v} \in V(G)$. Since this sampler is defined on the joint space $R$ and $V(G)$, we refer to it as *joint-space MCMC sampler*. Given a state $\mathfrak{s}$ of the chain, we denote by $\mathfrak{s}.\mathfrak{r}$ the first element of $\mathfrak{s}$, which is a vertex in $R$; and by $\mathfrak{s}.\mathfrak{v}$ the second element of $\mathfrak{s}$, which is a vertex in $V(G)$.

Our joint-space MCMC sampler consists of the following steps:

- First, we choose a pair $\langle \mathfrak{r}_0, \mathfrak{v}_0 \rangle$, as the initial state, where $\mathfrak{r}_0$ and $\mathfrak{v}_0$ are chosen uniformly at random from $R$ and $V(G)$, respectively.
- Then, at each iteration $t$, $1 \le t \le T$:
  - Let $\mathfrak{s}(t)$ be the current state of the chain.
  - We choose elements $\mathfrak{r}(t) \in R$ and $\mathfrak{v}(t) \in V(G)$, uniformly at random.
  - With probability $\min\left\{1, \frac{\delta_{\mathfrak{v}(t)\bullet}(\mathfrak{r}(t))}{\delta_{\mathfrak{s}(t).\mathfrak{v}\bullet}(\mathfrak{s}(t).\mathfrak{r})}\right\}$ we move from state $\mathfrak{s}(t)$ to the state $\langle \mathfrak{r}(t), \mathfrak{v}(t) \rangle$.

Techniques similar to our joint-space MCMC sampler have been used in *statistical physics* to estimate *free energy differences* [3]. Our joint-space MCMC sampler is a Metropolis-Hastings algorithm that possesses a *unique stationary distribution* [15, 20] defined as follows:

$$\mathbb{P}[r, v] = \frac{\delta_{v\bullet}(r)}{\sum_{r' \in R} \sum_{v' \in V(G)} \delta_{v'\bullet}(r')}. \quad (5)$$

All samples that have a specific value $r$ for their $\mathfrak{r}$ component form an Independence Metropolis-Hastings chain that possesses the stationary distribution defined in Equation 1. Samples drawn by our MCMC and joint-space MCMC samplers are non-iid. In Theorem 4.2, we show how our joint-space MCMC sampler can be used to estimate the ratios of betweenness scores of the vertices in $R$.

Theorem 4.2. *In our joint-space MCMC sampler, for any two vertices $r_i, r_j \in R$, we have:*

$$\frac{BC(r_i)}{BC(r_j)} = \frac{\mathbb{E}_{\mathbb{P}_{r_j}[v]}\left[\min\left\{1, \frac{\delta_{v\bullet}(r_i)}{\delta_{v\bullet}(r_j)}\right\}\right]}{\mathbb{E}_{\mathbb{P}_{r_i}[v]}\left[\min\left\{1, \frac{\delta_{v\bullet}(r_j)}{\delta_{v\bullet}(r_i)}\right\}\right]} \quad (6)$$

*where $\mathbb{E}_{\mathbb{P}_{r_i}[v]}$ (respectively $\mathbb{E}_{\mathbb{P}_{r_j}[v]}$) denotes the expected value with respect to $\mathbb{P}_{r_i}[v]$ (respectively $\mathbb{P}_{r_j}[v]$).*

The proof of Theorem 4.2 is based on the *detailed balance property* of Metropolis-Hastings algorithms and can be found in [10].

Let $r_i, r_j \in R$, and $M(i)$ and $M(j)$ be the multi-sets of samples taken by our joint-space MCMC sampler whose $\mathfrak{r}$ components are respectively $r_i$ and $r_j$. Equation 6 suggests to estimate $\frac{BC(r_i)}{BC(r_j)}$ as the ratio:

$$\frac{\frac{1}{|M(j)|} \times \sum_{\mathfrak{s} \in M(j)} \min\left\{1, \frac{\delta_{\mathfrak{s}.\mathfrak{v}}(r_i)}{\delta_{\mathfrak{s}.\mathfrak{v}}(r_j)}\right\}}{\frac{1}{|M(i)|} \times \sum_{\mathfrak{s} \in M(i)} \min\left\{1, \frac{\delta_{\mathfrak{s}.\mathfrak{v}}(r_j)}{\delta_{\mathfrak{s}.\mathfrak{v}}(r_i)}\right\}}. \tag{7}$$

We use Equation 7 to estimate the *ratio of the betweenness scores* of $r_i$ and $r_j$. We then define the *relative betweenness score* of $r_i$ with respect to $r_j$, denoted by $BC_{r_j}(r_i)$, as follows:

$$BC_{r_j}(r_i) = \frac{1}{|V(G)|} \sum_{v \in V(G)} \min\left\{1, \frac{\delta_{v\bullet}(r_i)}{\delta_{v\bullet}(r_j)}\right\}. \tag{8}$$

When we want to compare betweenness centrality of vertices $r_i$ and $r_j$, using *relative betweenness score* makes more sense than using the *ratio of betweenness scores*. In relative betweenness centrality, for each $v \in V(G)$, the ratio of the dependency scores of $v$ on $r_i$ and $r_j$ is computed and in the end, all the ratios are summed. Hence, for each vertex $v$ independent from the others, the effects of $r_i$ and $r_j$ on the shortest paths starting from $v$ are examined. Note that the notion of *relative betweenness score* can be further extended and presented as follows:

$$BC_{r_j}(r_i) = \frac{\sum_{v \in V(G)} \sum_{t \in V(G) \setminus \{v\}} \min\left\{1, \frac{\delta_{vt}(r_i)}{\delta_{vt}(r_j)}\right\}}{|V(G)| \cdot (|V(G)| - 1)}.$$

In the following, we show that the numerator of Equation 7, i.e.,

$$\frac{1}{|M(j)|} \sum_{\mathfrak{s} \in M(j)} \min\left\{1, \frac{\delta_{\mathfrak{s}.\mathfrak{v}}(r_i)}{\delta_{\mathfrak{s}.\mathfrak{v}}(r_j)}\right\},$$

can accurately estimate $BC_{r_j}(r_i)$. We refer to this value as $\ddot{BC}_{r_j}(r_i)$. For a pair of vertices $r_i, r_j \in R$, in Theorem 4.3 we derive an error bound for $\ddot{BC}_{r_j}(r_i)$.

Theorem 4.3. *Let $r_i, r_j \in R$, $M(j)$ be the multi-set of samples whose $\mathfrak{r}$ components are $r_j$, and $\overline{\delta(r_j)}$ be the average of dependency scores of vertices in $V(G)$ on $r_j$, i.e., $\overline{\delta(r_j)} = \frac{\sum_{v \in V(G)} \delta_{v\bullet}(r_j)}{|V(G)|}$. Suppose that there exists some value $\mu(r_j)$ such that for each vertex $v \in V(G)$, the following holds: $\delta_{v\bullet}(r_j) \leq \mu(r_j) \times \overline{\delta(r_j)}$. Then, for a given $\epsilon \in \mathbb{R}^+$, by our joint-space MCMC sampler and starting from any arbitrary initial state, we have*

$$\mathbb{P}\left[|\ddot{BC}_{r_j}(r_i) - BC_{r_j}(r_i)| > \epsilon\right] \leq 2\exp\left\{-\frac{|M(j)| - 1}{2}\left(\frac{2\epsilon}{\mu(r_i)} - \frac{3}{|M(j)| - 1}\right)^2\right\}. \tag{9}$$

Similar to Inequality 3, Inequality 9 does not depend on the initial state and it holds without need for burn-in. Furthermore, for given values $\epsilon \in \mathbb{R}^+$ and $\delta \in (0, 1)$, if we have

$$|M(j)| \geq \frac{\mu(r_j)^2}{2\epsilon^2} \ln \frac{2}{\delta},$$

then our joint-space MCMC sampler can estimate relative betweenness score of $r_i$ with respect to $r_j$ within an additive error $\epsilon$ with a probability at least $1 - \delta$.

## 5 CONCLUSION

In this paper, first given a network $G$ and a vertex $r \in V(G)$, we proposed a Metropolis-Hastings MCMC algorithm that samples from the space $V(G)$ and estimates betweenness score of $r$. We showed that our MCMC sampler provides an $(\epsilon, \delta)$-approximation. Then, given a network $G$ and a set $R \subset V(G)$, we presented a Metropolis-Hastings MCMC sampler that samples from the joint space $R$ and $V(G)$ and estimates relative betweenness scores of the vertices in $R$. We showed that for any pair $r_i, r_j \in R$, the ratio of the expected values of the estimated relative betweenness scores of $r_i$ and $r_j$ with respect to each other is equal to the ratio of their betweenness scores. We also showed that our joint-space MCMC sampler provides an $(\epsilon, \delta)$-approximation of the relative betweenness score of $r_i$ with respect to $r_j$. We leave efficient implementations of our proposed algorithms and evaluating their empirical efficiency for future work.

## ACKNOWLEDGMENTS

This work has been funded by the ANR project IDOLE.

## REFERENCES

[1] Manas Agarwal, Rishi Ranjan Singh, Shubham Chaudhary, and Sudarshan Iyengar. 2014. Betweenness Ordering Problem : An Efficient Non-Uniform Sampling Technique for Large Graphs. *CoRR* abs/1409.6470 (2014). http://arxiv.org/abs/1409.6470
[2] D. A. Bader, S. Kintali, K. Madduri, and M. Mihail. 2007. Approximating betweenness centrality. In *WAW*. 124–137.
[3] C. H. Bennett. 1976. Efficient estimation of free energy differences from Monte-Carlo data. *J. Comput. Phys.* 22 (1976), 245.
[4] Michele Borassi and Emanuele Natale. 2016. KADABRA is an ADaptive Algorithm for Betweenness via Random Approximation. In *ESA*. 20:1–20:18.
[5] U. Brandes. 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology* 25, 2 (2001), 163–177.
[6] U. Brandes and C. Pich. 2007. Centrality estimation in large networks. *Intl. Journal of Bifurcation and Chaos* 17, 7 (2007), 303–318.
[7] Ümit V. Çatalyürek, Kamer Kaya, Ahmet Erdem Sariyüce, and Erik Saule. 2013. Shattering and Compressing Networks for Betweenness Centrality. In *Proceedings of the 13th SIAM International Conference on Data Mining*. 686–694.
[8] Mostafa Haghir Chehreghani. 2014. Effective co-betweenness centrality computation. In *Seventh ACM International Conference on Web Search and Data Mining*. 423–432.
[9] Mostafa Haghir Chehreghani. 2014. An Efficient Algorithm for Approximate Betweenness Centrality Computation. *Comput. J.* 57, 9 (2014), 1371–1382.
[10] Mostafa Haghir Chehreghani, Talel Abdessalem, and Albert Bifet. 2017. Metropolis-Hastings Algorithms for Estimating Betweenness Centrality in Large Networks. *CoRR* abs/1704.07351 (2017). arXiv:1704.07351 http://arxiv.org/abs/1704.07351
[11] Mostafa Haghir Chehreghani, Albert Bifet, and Talel Abdessalem. 2018. Efficient Exact and Approximate Algorithms for Computing Betweenness Centrality in Directed Graphs. In *Advances in Knowledge Discovery and Data Mining (PAKDD)*. 752–764.
[12] Elizabeth M. Daly and Mads Haahr. [n. d.]. Social Network Analysis for Information Flow in Disconnected Delay-Tolerant MANETs. *IEEE Trans. Mob. Comput.* 8, 5 ([n. d.]), 606–621.
[13] M. Everett and S. Borgatti. 1999. The centrality of groups and classes. *Journal of Mathematical Sociology* 23, 3 (1999), 181–201.
[14] L. C. Freeman. 1977. A set of measures of centrality based upon betweenness, Sociometry. *Social Networks* 40 (1977), 35–41.
[15] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. 1996 (ISBN: 0-412-05551-1). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.
[16] M. Girvan and M. E. J. Newman. 2002. Community structure in social and biological networks. *Natl. Acad. Sci. USA* 99 (2002), 7821–7826.
[17] Krzysztof Łatuszyński, Błażej Miasojedow, and Wojciech Niemiro. 2012. *Nonasymptotic Bounds on the Mean Square Error for MCMC Estimates via Renewal Techniques*. Springer Berlin Heidelberg, Berlin, Heidelberg, 539–555.
[18] K. L. Mengersen and R. L. Tweedie. 1996. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* 24, 1 (02 1996), 101–121.
[19] K. L. Mengersen and R. L. Tweedie. 1996. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics* 24, 1 (Feb. 1996), 101–121.
[20] S. P. Meyn and R. L. Tweedie. 1993. *Markov chains and stochastic stability*. Springer-Verlag, London.
[21] Matteo Riondato and Evgenios M. Kornaropoulos. 2016. Fast approximation of betweenness centrality through sampling. *Data Mining and Knowledge Discovery* 30, 2 (2016), 438–475.
[22] Matteo Riondato and Eli Upfal. 2016. ABRA: Approximating Betweenness Centrality in Static and Dynamic Graphs with Rademacher Averages. In *KDD*. 1145–1154.
[23] Y. Wang, Z. Di, and Y. Fan. 2011. Identifying and Characterizing Nodes Important to Community Structure Using the Spectrum of the Graph. *PLoS ONE* 6, 11 (2011), e27418.