An Approximation Algorithm for the Rich-Club Coefficient via Sample Complexity

Eric Y. Fukuyama[1], **Leandro M. Zatesko**[1], **Murilo V. G. da Silva**[2]

[1]Federal University of Technology of Paraná – Curitiba (UTFPR-CT), Brazil

[2]Federal University of Paraná (UFPR), Brazil

ericfukuyama@utfpr.alunos.edu.br, zatesko@utfpr.edu.br, murilo@inf.ufpr.br

Abstract. *Sample complexity theory has been used to design efficient approximation algorithms for computing graph parameters, such as betweenness centrality and clustering coefficient. The Rich-Club coefficient $\phi(k)$ quantifies the density of edges among vertices with degree greater than k. Using sample complexity theory, we present a (p, ε)-relative approximation algorithm for the problem of determining $\phi(k)$ for all k. That is, given parameters $0 < p, \varepsilon, \delta < 1$, our algorithm returns, with probability at least $1 - \delta$, an estimate $\hat{\phi}(k)$ with relative error at most ε for every k such that $\phi(k) \geq \sigma_k(p)$, where $\sigma_k(p)$ denotes a threshold function. The algorithm runs in time $\mathcal{O}\Big(\frac{\Delta}{\varepsilon^2 p}\Big(\log\Delta\log\frac{1}{p} + \log\frac{1}{\delta}\Big) + n + m\Big)$, improving over the exact approach, which requires $\mathcal{O}(\Delta m)$ time.*

1. Introduction

Throughout the text, we follow the notation and terminology adopted in the literature, especially in [Har-Peled 2011, Riondato and Kornaropoulos 2016].

The Rich-Club coefficient is a graph-theoretic metric that quantifies the density of edges among high-degree vertices. The idea of the coefficient was introduced by [Zhou and Mondragon 2004]. Moreover, [Colizza et al. 2006] defines the Rich-Club coefficient as $\phi(k) = E_{>k}/\binom{N_{>k}}{2}$, where $N_{>k}$ is the cardinality of the set of vertices with degree strictly greater than k and $E_{>k}$ is the cardinality of the set of edges with both endpoints in $N_{>k}$. Recall that $\binom{N_{>k}}{2}$ is the total number of distinct pairs that can be formed from the $N_{>k}$ vertices.

The problem addressed in this work is, given a graph $G = (V, E)$, to compute the value of the Rich-Club coefficient $\phi(k)$ for every k in the interval $[0, \Delta]$. In the exact approach, computing $\phi(k)$ requires, for each value of k, counting the edges that connect vertices whose degree is greater than k. This operation may require scanning a large number of edges, proportional to $m = |E|$, for each of the Δ values of k, yielding an overall $O(\Delta m)$ complexity. Moreover, no exact algorithm with asymptotically better running time is known, to the best of our knowledge. On large-scale graphs, this cost can become prohibitive.

The use of approximation algorithms to compute graph metrics is not new and has been widely investigated. The motivation is clear: many fundamental metrics, such as centralities or clustering coefficients, have high exact computational cost in large-scale settings. Hence, sampling methods, probabilistic techniques, and complexity bounds based on the VC-dimension have proved to be viable alternatives to make

computation more scalable. Among notable examples are approximation algorithms for centrality measures such as betweenness centrality, the Percolation Centrality Problem, the All-Pairs Shortest Paths (APSP) problem, and the computation of the local clustering coefficient. In all these cases, methods based on statistical sampling have been proposed, drastically reducing running time in exchange for a controlled error margin [de Lima 2022, Riondato and Kornaropoulos 2016].

To make the presentation self-contained, we now recall some concepts from sample complexity theory that will be used in the next sections. Definitions 1, 2, and 3 can be found in [Har-Peled 2011]. Furthermore, Definition 4 and Theorem 1 are from [Riondato and Kornaropoulos 2016].

Definition 1 (Range Space)**.** A *range space* is a pair $A = (X, \mathcal{I})$, where X is a set and $\mathcal{I}$ is a family of subsets of X. The elements of X are called *points*, and the elements of $\mathcal{I}$ are called *ranges*. For $S \subseteq X$, we define $\mathcal{I}_S = \{S \cap I : I \in \mathcal{I}\}$, which is the projection (or restriction) of $\mathcal{I}$ onto S.

Definition 2 (Shattered Set)**.** If $\mathcal{I}_S$ contains all subsets of S (that is, if $\mathcal{I}_S = 2^S$; in particular, if S is finite, then $|\mathcal{I}_S| = 2^{|S|}$), we say that S is *shattered* by $\mathcal{I}$.

Definition 3 (VC-Dimension)**.** The Vapnik–Chervonenkis dimension (VC-dimension) of a *range space* $A = (X, \mathcal{I})$, denoted $\mathrm{VCDim}(A)$, is the maximum cardinality of a subset of X that is shattered by $\mathcal{I}$. If there exist shattered subsets of arbitrarily large cardinality, then $\mathrm{VCDim}(A) = \infty$.

Definition 4 (Relative (p, ε)-approximation)**.** Let R be a range set on X and π be a probability distribution on X. For $p, \varepsilon \in (0, 1)$, a relative (p, ε)-approximation to (R, π) is a bag S of elements from X such that for any $A \in R$ such that $\pi(A) \geq p$, we have $|\pi(A) - \pi_S(A)| \leq \varepsilon\pi(A)$, and for any $B \in R$ such that $\pi(B) < p$, we have $\pi_S(B) \leq (1 + \varepsilon)p$.

Theorem 1. *Let R be a range set on a domain X with $\mathrm{VCDim}(R) \leq d$, and let π be a distribution on X. Given $\varepsilon, \delta, p \in (0, 1)$, let S be a collection of $|S|$ points from X sampled according to π, with $|S| \geq \frac{c'}{\varepsilon^2 p}\left(d \log \frac{1}{p} + \log \frac{1}{\delta}\right)$, where c' is an absolute positive constant. Then S is a relative (p, ε)-approximation to (R, π) with probability at least $1 - \delta$.*

Where c' is empirically estimated to be 0.5.

2. Results

Algorithm 1 takes as input a graph G, a maximum tolerated error ε, a confidence level δ, and a parameter p related to the threshold above which we obtain a good approximation. Moreover, the constant c that appears in the algorithm is the same universal constant from Theorem 1.

To represent the degree vector in the algorithm, we use the notation $\deg[i]$, where i is the index of the i-th vertex of the graph. The symbol Δ denotes the maximum degree of the graph. The number of edges for which the minimum degree among the two endpoints is greater than i is denoted by $E_{>i}$. Analogously, $V_{>i}$ denotes the number of vertices whose degree is greater than i. The neighborhood size of a is denoted by N_a. We denote by $\hat{\rho}$ an estimate of ρ. Finally, the output of the algorithm is the vector of estimates $\hat{\phi}$, of

Algorithm 1 CoefficientRichClubEstimation$(G, \varepsilon, \delta, p)$

Procedure CoefficientRichClubEstimation($G = (V, E)$, ε, δ, p)
Input: Graph $G = (V, E)$ with m edges; parameters $0 < \varepsilon, \delta, p < 1$.
Output: An estimate of the Rich-Club coefficient $\{\hat{\phi}(k)\,;\, k \in \mathbb{N},\, k \leq \Delta\}$.
for $i \leftarrow 1$ **to** n **do**
    $\deg[i] \leftarrow |N_{v_i}|$
end for
$\Delta \leftarrow \max\big(\{\deg[i]\,;\, 1 \leq i \leq n\}\big)$
for $i \leftarrow 1$ **to** Δ **do**
    $\text{freqDeg}[i] \leftarrow 0$
end for
for $i \leftarrow 1$ **to** n **do**
    $\text{freqDeg}[\deg[i]] \leftarrow \text{freqDeg}[\deg[i]] + 1$
end for
$V_{>\Delta} \leftarrow 0$
for $i \leftarrow \Delta - 1$ **downto** 1 **do**
    $V_{>i} \leftarrow V_{>i+1} + \text{freqDeg}[i+1]$
end for
$r \leftarrow \left\lceil \frac{c}{\varepsilon^2 p}\left(\left(\lfloor \lg(\Delta - 1) \rfloor + 1\right) \cdot \log \frac{1}{p} + \log \frac{1}{\delta}\right)\right\rceil$
for $i \leftarrow 1$ **to** Δ **do**
    $\hat{E}_{>i} \leftarrow 0$
end for
for $t \leftarrow 1$ **to** r **do**
    sample an edge $e = \{a, b\} \in E$ uniformly at random
    $k \leftarrow \min\{\deg[a], \deg[b]\}$
    for $i \leftarrow 1$ **to** $k - 1$ **do**
        $\hat{E}_{>i} \leftarrow \hat{E}_{>i} + \frac{m}{r}$
    end for
end for
for $i \leftarrow 1$ **to** $\Delta - 1$ **do**
    $\hat{\phi}[i] \leftarrow \frac{2\,\hat{E}_{>i}}{V_{>i}(V_{>i} - 1)}$
end for
End Procedure return $\hat{\phi}$

size Δ, containing the estimated values of the Rich-Club coefficient for values of k such that $\phi(k) \geq \sigma_k(p)$, where $\sigma_k(p) = \frac{pm}{\binom{N_{\geq k}}{2}}$.

We first prove Theorem 2, which is inspired by a theorem in [de Lima 2022] and plays a fundamental role in the proof of Theorem 3. Then, we present Theorem 4, which establishes the running-time complexity of the algorithm. In approximation algorithms based on sample complexity, it is essential to define the number of samples required to guarantee given parameters. In this context, finding an upper bound on the VC-dimension is directly related to establishing a lower bound on the sample size.

Theorem 2. $\text{VCDim}(R) \leq \lfloor \lg(\Delta - 1) \rfloor + 1$, *where* $R = (X, \mathcal{I})$ *is the range space*

in which X is the edge set E and $\mathcal{I}$ is the collection of sets I_k, with $I_k = \{e \in E :$ the two endpoints of e have degree greater than $k\}$.

Proof. An edge $e \in E$ belongs to an interval I_k if its two endpoints have degree greater than k. Let g be the minimum degree among the two endpoints of e. Thus, the maximum number of sets in $\mathcal{I}$ that can contain e is $\Delta - 1$. Assume that $\mathrm{VCDim}(R) = d$. Then e must appear in 2^{d-1} intervals, because, by Definition 3, we have a set S with $|S| = d$ and $|\mathcal{I}_S| = 2^d$, where $\mathcal{I}_S = \{I_k \cap S\,;\, I_k \in \mathcal{I}\}$. Hence, each intersection $I_k \cap S$ must represent a subset of S, so edge e must appear in half of the intervals. Therefore, $2^{d-1} \leq \Delta - 1 \implies d \leq \lfloor \lg(\Delta - 1) \rfloor + 1$. □

Theorem 3. *Let c be the constant of Theorem 1. Given a graph $G = (V, E)$, let $S \subseteq E$ be a sample of size $r = \left\lceil \frac{c}{\varepsilon^2 p} \left((\lfloor \lg(\Delta - 1) \rfloor + 1) \log \frac{1}{p} + \log \frac{1}{\delta} \right) \right\rceil$, for constants $0 < p, \varepsilon, \delta < 1$. Algorithm 1 returns, with probability at least $1 - \delta$, an approximation $\hat{\phi}(k)$ of $\phi(k)$ with relative error ε, for each $k \in \{1, \ldots, \Delta\}$ such that $\phi(k) \geq \sigma_k(p)$, where $\sigma_k(p) = \frac{pm}{\binom{N_{>k}}{2}}$.*

Proof. First, by Theorem 1, for given values $0 < \varepsilon, \delta, p < 1$, a distribution π over X, and a universal constant $c > 0$, and using Theorem 2 as a bound on VCDim, we obtain $|S| \geq \frac{c}{\varepsilon^2 p} \left((\lfloor \lg(\Delta - 1) \rfloor + 1) \log \frac{1}{p} + \log \frac{1}{\delta} \right)$. In particular, $r = \left\lceil \frac{c}{\varepsilon^2 p} \left((\lfloor \lg(\Delta - 1) \rfloor + 1) \log \frac{1}{p} + \log \frac{1}{\delta} \right) \right\rceil$. Moreover, by Theorem 1, the sample of size r is a (p, ε)-relative approximation with probability at least $1 - \delta$.

For the rest of the proof, let α_k be the set of edges whose two endpoints have degree greater than k. For each $k \in \{1, \ldots, \Delta\}$, let $\mathbf{1}_k(e)$ be the indicator function that equals 1 if $e \in \alpha_k$ and 0 otherwise. Thus, $E_{>k} = \sum_{e \in E} \mathbf{1}_k(e)$. The estimate $\hat{E}_{>k}$ computed by the algorithm is incremented by m/r whenever an edge $e \in S$ belongs to α_k, i.e., $\hat{E}_{>k} = \sum_{e \in S} \frac{m}{r} \mathbf{1}_k(e)$. Note that $\hat{E}_{>k} = \sum_{e \in S} \frac{m}{r} \mathbf{1}_k(e) = m \frac{|S \cap \alpha_k|}{|S|}$. Since S is a (p, ε)-relative approximation, we have $\frac{|E_{>k} - \hat{E}_{>k}|}{E_{>k}} = \frac{\left| \Pr_\pi(\alpha_k) - \frac{|S \cap \alpha_k|}{|S|} \right|}{\Pr_\pi(\alpha_k)} \leq \varepsilon$. Equivalently, $|E_{>k} - \hat{E}_{>k}| \leq \varepsilon E_{>k}$. Hence, $|\phi(k) - \hat{\phi}(k)| = \frac{2|E_{>k} - \hat{E}_{>k}|}{N_{>k}(N_{>k} - 1)} \leq \frac{2\varepsilon E_{>k}}{N_{>k}(N_{>k} - 1)} = \varepsilon \phi(k)$.

Combining this result with Theorem 1 and 2, and using a sample S of size $r = \left\lceil \frac{c}{\varepsilon^2 p} \left((\lfloor \lg(\Delta - 1) \rfloor + 1) \log \frac{1}{p} + \log \frac{1}{\delta} \right) \right\rceil$, we guarantee that Algorithm 1 provides a (p, ε)-approximate estimate of $\phi(k)$ with probability at least $1 - \delta$, for every $k \in [1, \Delta]$ such that $\phi(k) \geq \sigma_k(p)$. Indeed, $\Pr_\pi(\alpha_k) \geq p$ if and only if $\phi(k) \geq \sigma_k(p)$, since $\phi(k) = \frac{E_{>k}}{\binom{N_{>k}}{2}} = \frac{m \Pr_\pi(\alpha_k)}{\binom{N_{>k}}{2}}$. □

Theorem 4. *Algorithm 1 runs in time $\mathcal{O}(r\Delta + n + m)$, and, since $r = \mathcal{O}\left(\frac{1}{\varepsilon^2 p} \left(\log \Delta \log \frac{1}{p} + \log \frac{1}{\delta} \right) \right)$, the total running time of Algorithm 1 is $\mathcal{O}\left(\frac{\Delta}{\varepsilon^2 p} \left(\log \Delta \log \frac{1}{p} + \log \frac{1}{\delta} \right) + n + m \right)$* □.

References

Colizza, V., Flammini, A., Serrano, M. A., and Vespignani, A. (2006). Detecting rich-club ordering in complex networks. *Nature Physics*, 2(2):110–115.

de Lima, A. M. (2022). *Approximation Algorithms in Graphs via Sample Complexity*. PhD thesis, Federal University of Paraná (UFPR), Curitiba. PhD thesis (Doctorate in Computer Science) – Exact Sciences Division, Graduate Program in Computer Science.

Har-Peled, S. (2011). *Geometric Approximation Algorithms*. American Mathematical Society, USA. Monographs on Discrete Mathematics and Applications.

Riondato, M. and Kornaropoulos, E. M. (2016). Fast approximation of betweenness centrality through sampling. *Data Mining and Knowledge Discovery*, 30(2):438–475.

Zhou, S. and Mondragon, R. (2004). The rich-club phenomenon in the internet topology. *IEEE Communications Letters*, 8(3):180–182.