A Linear-Time Algorithm and Analysis of Graph Relative Hausdorff Distance*

Sinan G. Aksoy†, Kathleen Nowak†, and Stephen J. Young†

Abstract. Graph similarity metrics serve far-ranging purposes across many domains in data science. As graph datasets grow in size, scientists need comparative tools that capture meaningful differences, yet are lightweight and scalable. Graph Relative Hausdorff (RH) distance is a promising, recently proposed measure for quantifying degree distribution similarity. In spite of recent interest in RH distance, little is known about its properties. Here, we conduct an algorithmic and analytic study of RH distance. In particular, we provide the first linear-time algorithm for computing RH distance, analyze examples of RH distance between families of graphs, and prove several analytic results concerning the range, density, and extremal behavior of RH distance values.

Key words. Relative Hausdorff distance, degree distribution, graph similarity measure, network science

AMS subject classifications. 05C82, 05C07, 05C85

DOI. 10.1137/19M1248224

1. Introduction and background. The need to quantify the similarity between graphs arises in domains including graph sampling techniques [22, 24, 36], generative model selection [19, 28], botnet discovery [37], anomaly detection [21, 31], classification and clustering of graph ensembles [1, 12], and epidemic spreading [27, 30]. In choosing a similarity function, the data scientist has no dearth of options at her disposal. Graph edit distance [32], maximum common subgraph-based distance [17], iterative vertex-neighborhood methods [8, 20], or indeed the difference in any selection of structural metrics like graph diameter, clustering coefficients, modularity, centrality, or eigenvalues, could potentially serve as similarity measures. However, as graph data becomes more massive, varied, and dense, there is an increasing need for comparison tools that are not only informative but *lightweight* and *scalable*. In this regard, a number of tools unfortunately fall short. For example, while the aforementioned edit distance provides a nuanced and intuitive measure of similarity, its exact computation is NP-complete, and even its approximation is APX-hard [25].

One quantity that is easy to extract from data, and often used when comparing graphs, is the degree distribution. Perhaps the most fundamental and oft-studied notion in network science, the degree distribution plays an important role in graph modeling, algorithms, and resilience [6, 11, 29]. Due partly to its tractability and heavy-tailed nature, the degree distribution is particularly important for analyzing massive graphs like the Internet [9, 16]. Notwithstanding its ubiquity and importance, the degree distribution has limitations in capturing network structure. For example, it is straightforward to construct ensembles of graphs with identical degree distributions but strikingly different structural properties [13]. Nonetheless,

because of its practicality, accessibility, and interpretability, the degree distribution remains a key representation of interest in graph data analysis.

Despite the apparent popularity of degree distributions, there is little consensus on how to meaningfully quantify their similarity. Although elegant and succinct, closed-form characterizations like the famous power law are prone to statistical error when fitted to empirical data [10]. A number of general-purpose statistical tools for probability distributions, such as the Kolmogorov–Smirnov (KS) test statistic, Earth mover's distance, or percentile methods [19], avoid imposing assumptions on degree distribution shape but, nonetheless, lack stringency in the graph setting [4, 26, 33]. Last but not least, researchers have proposed feature extraction methods which, based on the standard deviation of the degree distribution, extract a fixed-length feature vector from the degree distribution which is then used to define a distance function between distributions [4]. While a single, definitive metric for degree distribution comparison is clearly unrealistic, we believe a detailed, analytic study of the metrics themselves provides utility and garners insight into what the different metrics capture.

In this work, we conduct an algorithmic and analytical study of a newly proposed, sophisticated measure of degree distribution similarity called *Relative Hausdorff (RH) distance*. Graph RH distance was first introduced by Simpson, Seshadhri, and McGregor [33] in the context of *small-space streaming algorithms*. Given a sequence or stream of edges $e_1, \ldots, e_m$ of a graph, they devised an algorithm to estimate the graph's degree distribution when only a small fraction (typically less than 1%) of the edge stream can be stored in memory at a time before being forgotten. In order to score the performance of this algorithm, they sought a similarity measure by which to quantitatively compare two graph degree distributions. Unsatisfied with the flaws in existing metrics commonly used for comparing degree distributions, Simpson et al. propose a new measure inspired by the Hausdorff metric from topology [18], aimed at capturing degree distribution closeness at all scales.

In the context of degree distribution comparison, the purported benefits of RH distance over other metrics stem from a combination of factors. First, RH distance is flexible in tolerating variable error in both vertex degrees and their relative frequencies. This is especially useful for comparing heavy-tailed distributions typical of real graphs: for low degrees (when frequencies are large), RH distance allows for some error in frequency but less so in degree values, whereas for high degrees (when frequencies are small), RH distance allows more error in degree estimates but less in frequency. Second, while flexible in this regard, RH distance is also stringent in requiring closeness at *every* point of the head, tail, and middle of degree distributions. Accordingly, any outlier behavior in one distribution must be approximated in the other. In contrast, a number of existing metrics like KS distance are insensitive to differences in the tails of degree distributions [33]. For these reasons, RH distance appears a promising tool for graph degree distribution comparison. Yet despite a spate of subsequent work [3, 14, 26, 35] applying RH distance to graph data, as well as recent work by the authors analyzing RH distance as an anomaly-detection method in cybersecurity applications [2], little is known about the fundamental properties of this metric.

We organize our work as follows: in section 2, we define the original RH metric, as well as related variants from the literature. In section 3, improving upon a previously known quadratic algorithm, we propose a new algorithm for computing RH distance and show its time complexity is linear in the maximum degrees of the graphs being compared. This algorithm

is not only of practical interest but also provides a blueprint for analytically determining the RH distance between families of graphs. In section 4, we obtain closed-form solutions for the RH distance between common families of graphs; these examples demonstrate that, contrary to its description in the literature as a "distance metric," RH distance does not satisfy the triangle inequality. In section 5 we analyze the possible values one can attain when computing the RH distance between two graphs. In particular, we (1) determine the *range* of possible RH values by establishing sharp upper bounds on the maximum RH distance between graphs, (2) show that the RH distance is *dense* in the real numbers, and (3) characterize the *extremal* behavior of the RH distance between graphs on a fixed vertex set that differ by a single edge. Throughout, we compare the theoretical results against those obtained by calculating the RH distance on pairs of real-world graphs from the Stanford Network Analysis Project (SNAP) [23]. This collection contains 40 graphs[1] with a variety of different structural properties and sizes. Finally, in section 6 we outline open questions on RH distance and briefly suggest avenues for future work.

2. Preliminaries. The RH distance between graphs is based on their complementary cumulative degree histograms (ccdhs), which summarize their vertex degree counts. As the ccdh is closely related to the oft-mentioned degree sequence and degree distribution of graphs, we define all three below for clarity.

Definition 2.1. *For an n-vertex graph $G = (V, E)$ with no isolated vertices, let $d(v)$ denote the degree of a vertex v.*
  (a) *The degree sequence of G is $(d(v_i))_{i=1}^{n}$.*
  (b) *The degree distribution (i.e., degree histogram) of G is $(n(k))_{k=1}^{\infty}$, where $n(k) = |\{v \in V : d(v) = k\}|$.*
  (c) *The complementary cumulative degree histogram (ccdh) of G is $(N(k))_{k=1}^{\infty}$, where $N(k) = \sum_{i=k}^{\infty} n(i)$.*

Up to vertex labeling, all three of the above notions are equivalent in the sense that, for a given graph, each can be uniquely obtained from any of the others. However, for ease of exposition, it is often preferable to summarize degree counts using the ccdh, which (as noted in [33]) is typically less noisy than the degree distribution and additionally has the nice property of being monotonically decreasing. When clear from context, we will slightly abuse notation and write $F(d)$ for a graph F to mean the value of the ccdh of F at d. Furthermore, we let $\Delta(F)$ denote the maximum degree of graph F. Below, we present the original definition of RH distance, using slightly different notation than in [33].

Definition 2.2 (discrete RH [33]). *Let F, G be graphs. The discrete directional RH distance from F to G, denoted $\overrightarrow{\mathcal{RH}}_d(F, G)$, is the minimum ϵ such that*

$$\forall d \in \{1, \ldots, \Delta(F)\}, \exists d' \in \{1, \ldots, \Delta(G)+1\} \text{ such that } |d-d'| \leq \epsilon d \text{ and } |F(d)-G(d')| \leq \epsilon F(d).$$

We call $\mathcal{RH}_d(F, G) = \max\{\overrightarrow{\mathcal{RH}}_d(F, G), \overrightarrow{\mathcal{RH}}_d(G, F)\}$ the discrete RH distance between F and G.

---

[1]All datasets were processed as simple graphs, ignoring any multiple, weighted, directed, or self-loop edges.

By definition, $\mathcal{RH}_d(F, G) = \epsilon$ means that for every degree k in the graph F, $F(k)$ is within ϵ-fractional error of some $G(k')$, where k' is within ϵ-fractional error of k. In this sense, the RH measure is flexible in accommodating some error in both vertex degree values as well as their respective counts, yet strict in requiring that *every point* in F be ϵ-close to G (and vice versa). Crucially, we note that the ccdh underlying Definition 2.2 is defined on a discrete domain.

In [26, 35], Matulef and Stolman argue that the discrete domain of the ccdh in the above definition of RH distance poses a potential flaw. In particular, they note that this may lead to counterintuitively large RH distances between pairs of graphs which are intuitively similar. Indeed, they give an example[2] of two graphs on the same vertex set which differ only by a single edge, yet still have (very large) RH distance of 1 for all values of n. To ameliorate this issue, they propose using a *smooth ccdh* in which each successive pair of points is connected via line segments. More formally, we define the smooth ccdh as follows.

Definition 2.3. *For a graph $G = (V, E)$, let $f(k) = |\{v : d(v) \geq k\}|$. The smooth ccdh of a graph G is a function $\varphi : \mathbb{Z}_{\geq 1} \times [0, 1) \to \mathbb{R}_{\geq 0}$ defined by*

$$\varphi(x, \epsilon) = (1 - \epsilon)f(x) + \epsilon f(x + 1).$$

For ease of notation, we may write $\varphi(x) := \varphi(\lfloor x \rfloor, x - \lfloor x \rfloor)$.

Using smooth ccdhs, the *smooth RH distance* is defined analogously as in Definition 2.2, except d' can now assume noninteger values.

Definition 2.4 (smooth RH [26, 35]). *Let F, G be graphs. The smooth directional RH distance from F to G, denoted $\overrightarrow{\mathcal{RH}}(F, G)$, is the minimum ϵ such that*

$$\forall d \in \{1, \dots, \Delta(F)\}, \exists d' \in [1, \Delta(G) + 1] \text{ such that } |d - d'| \leq \epsilon d \text{ and } |\varphi_F(d) - \varphi_G(d')| \leq \epsilon \varphi_F(d).$$

We call $\mathcal{RH}(F, G) = \max\{\overrightarrow{\mathcal{RH}}(F, G), \overrightarrow{\mathcal{RH}}(G, F)\}$ the smooth RH distance between F and G.

Comparing Definitions 2.2–2.4, it is clear that for any pair of graphs, their smooth RH distance is at most their discrete RH distance. Henceforth, we focus exclusively on smooth RH distance and thus will abuse notation and use $F(x)$ for $\varphi_F(x)$.

3. A linear-time algorithm for RH distance. The authors in [26, 33, 35] do not provide a pseudocode specification of an algorithm for computing (discrete) RH distance between graphs. However, to supplement [35], Python code implementing a straightforward, quadratic-time algorithm for computing smooth RH distance was made publicly available [34]. Here, we provide a linear-time algorithm (SMOOTHRH) for computing RH distance in graphs. As an aside, we note that a linear-time algorithm for computing (nongraph) Hausdorff distance between convex polygons was given in [5], although the same techniques do not apply in the case of graph RH distance.

For convenience of notation, for any graph H we will interpret the smooth ccdh as taking value 0 for any $x > \Delta(H) + 1$ even though this region is not in the domain. Before beginning

---

[2]Let G consist of the complete bipartite graph $K_{2,n-2}$ and G' be the same graph with one edge removed. By examining their respective ccdhs, it is clear that $|G'(n) - G(d)| \geq 1$ for every d; hence RH distance as given in Definition 2.2 is 1.

Algorithm 3.1 Maximum of ϵ and smooth RH distance from (x,y) to H

---

procedure SMOOTHRHDIST$((x,y), H, \epsilon)$

1: $(\ell_x, \ell_y) \leftarrow ((1-\epsilon)x, (1-\epsilon)y)$ $\triangleright$ lower left corner of box$((x,y);\epsilon)$

2: $(r_x, r_y) \leftarrow ((1+\epsilon)x, (1+\epsilon)y)$ $\triangleright$ upper right corner of box$((x,y);\epsilon)$

3: **if** $\ell_x < 1$ **and** $H(1) < \ell_y$ **then** $\triangleright$ H passes well above box$((x,y);\epsilon)$

4:    **return** $1 - \frac{H(1)}{y}$

5: **else if** $\ell_x \geq 1$ **and** $H(\ell_x) < \ell_y$ **then** $\triangleright$ H passes below box$((x,y);\epsilon)$

6:    $j \leftarrow \lceil \ell_x \rceil$

7:    **while** $H(j-1) < \frac{(j-1)y}{x}$ **and** $j > 1$ **do**

8:     $j \leftarrow j - 1$

9:    **end while**

10:   **if** $j = 1$ **then**

11:    **return** $1 - \frac{H(1)}{y}$

12:   **else**

13:    **return** $\frac{y+(x-j)H(j-1)-(x-j+1)H(j)}{y+x(H(j-1)-H(j))}$

14:   **end if**

15: **else if** $r_x < \Delta(H) + 1$ **and** $H(r_x) > r_y$ **then** $\triangleright$ H passes above box$((x,y);\epsilon)$

16:   $j \leftarrow \lfloor r_x \rfloor$

17:   **while** $H(j+1) > \frac{(j+1)y}{x}$ **do**

18:    $j \leftarrow j + 1$

19:   **end while**

20:   **return** $\frac{(j+1-x)H(j)+(x-j)H(j+1)-y}{y+x(H(j)-H(j+1))}$

21: **else** $\triangleright$ H passes through box$((x,y);\epsilon)$

22:   **return** ϵ

23: **end if**

Algorithm 3.2 Smoothed RH distance between the ccdhs F and G

---

procedure SMOOTHRH(F, G)

1: $\epsilon \leftarrow 0$

2: $\Delta \leftarrow \max\{\Delta(G), \Delta(F)\}$ $\triangleright$ maximum degree occurring in F or G

3: **for** $i \in 1, \ldots, \Delta$ **do**

4:    **if** $i \leq \Delta(F)$ **then**

5:     $\epsilon \leftarrow$ SMOOTHRHDIST$((i, F(i)), G, \epsilon)$

6:    **end if**

7:    **if** $i \leq \Delta(G)$ **then**

8:     $\epsilon \leftarrow$ SMOOTHRHDIST$(i, G(i)), F, \epsilon)$

9:    **end if**

10: **end for**

11: **return** ϵ

---

the analysis of the SMOOTHRHDIST and SMOOTHRH algorithms, we first introduce notation for the RH equivalent of a ball:

$$\text{box}((x,y);\epsilon) = \left\{ (s,t) \in \mathbb{R}^2 \colon |x-s| \leq \epsilon x, |y-t| \leq \epsilon y \right\}.$$

Thus, $\overrightarrow{\mathcal{RH}}(F,G)$ may be thought of as the least ϵ such that for all $1 \leq d \leq \Delta(F)$ the smooth ccdh for G intersects $\text{box}((d,F(d));\epsilon)$. We note that if $\epsilon < \epsilon'$, then $\text{box}((x,y);\epsilon) \subset \text{box}((x,y);\epsilon')$, and hence the correctness of SMOOTHRH follows immediately from the correctness of SMOOTHRHDIST.

To establish the correctness of SMOOTHRHDIST, we first note that as the smooth ccdhs are monotonically decreasing it is easy to determine whether the smooth ccdh for H intersects $\text{box}((x,y);\epsilon)$ for some fixed (x,y) and ϵ. Specifically, the smooth ccdh of H intersects $\text{box}((x,y);\epsilon)$ if and only if either

- $(1-\epsilon)x < 1$ and $H(1) \geq (1-\epsilon)y$, or
- the ccdh of H is defined at $(1-\epsilon)x$ and is greater than $(1-\epsilon)y$ and is undefined at $(1+\epsilon)x$, or
- the ccdh of H is defined at $(1-\epsilon)x$ and $(1+\epsilon)x$ and is greater than $(1-\epsilon)y$ at $(1-\epsilon)x$ and is less than $(1+\epsilon)y$ at $(1+\epsilon)x$.

It is straightforward to verify that the combined negation of the conditions on lines 3, 5, and 15 results in one of these conditions holding; thus $\text{box}((x,y);\epsilon)$ intersects the smooth ccdh of H and SMOOTHRHDIST correctly returns ϵ.

We now consider the situation when the smooth ccdh of H does not intersect any portion of $\text{box}((x,y);\epsilon)$. Broadly speaking, this situation can be defined into two cases; the ccdh is above the box or the ccdh is below the box. These scenarios are illustrated by the red (above the box) and green (below the box) lines in Figure 1(a).

In the case where $\text{box}((x,y);\epsilon)$ is below the smooth ccdh of H, the return value of SMOOTHRHDIST should be the least ϵ' such that $\mathbf{r} = ((1+\epsilon')x, (1+\epsilon')y)$ is on the smooth ccdh of H. Noting that $((1+\epsilon')x, (1+\epsilon')y)$ implicitly defines a line L and the smoothed ccdh
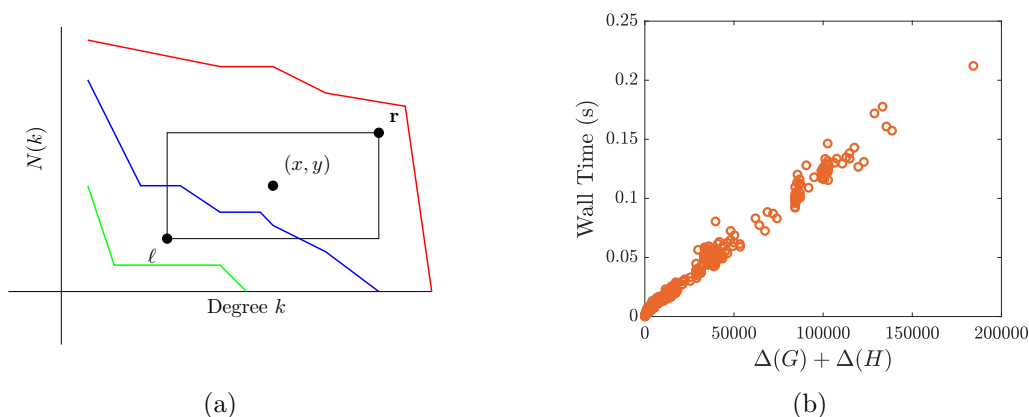

(a)	(b)

Figure 1. *An illustration of* $\text{box}((x,y);\epsilon)$ *together with ccdhs passing above (red), through (blue), and below (green) the box* (a). *Wall times for RH distance computation on the SNAP data* (b).

is piecewise linear, we can find such an ϵ' by identifying a pair $(j, j + 1)$ where L and the smooth ccdh for H cross. This is achieved by an iterative search over such pairs starting on line 17 and returning ϵ' which determines the unique point of intersection between L and the smooth ccdh of H.

For the case where $\text{box}((x, y); \epsilon)$ is above the smooth ccdh of H, there are essentially two cases: either the line L implicitly defined by $((1 - \epsilon')x, (1 - \epsilon')y)$ passes through the smooth ccdh of H or the desired ϵ' is given by when $(1 - \epsilon')y = H(1)$. Similarly as to the case where $\text{box}((x, y); \epsilon)$ is below the smooth ccdh of H, the intersection point with L is determined by iteratively searching to the left. The possibility that L doesn't intersect with the smooth ccdh is handled by the cases on lines 3 and 10. Together these three cases yield the following result.

Lemma 3.1. *Algorithm* SMOOTHRHDIST *correctly calculates the maximum of ϵ and the RH distance between the point (x, y) and the complimentary cumulative degree distribution H.*

We now turn to the overall run time of the algorithm SMOOTHRH.

Theorem 3.2. *The algorithm* SMOOTHRH *calculates the smooth RH distance between two complimentary cumulative degree distributions F and G in time at most $\Theta(\Delta(G) + \Delta(F))$.*

Proof. We note that a cursory examination of SMOOTHRH and SMOOTHRHDIST reveals that running time of SMOOTHRHDIST is driven by the number of times the while loops at lines 7 and 17 are executed. It is easy to see that this gives an $\mathcal{O}(\Delta(F)^2 + \Delta(G)^2)$ upper bound on the run time. To show that the run time is linear we will in fact show that these while loops are entered at most $\Delta(F) + \Delta(G)$ times giving a overall run time of $\mathcal{O}(\Delta(F) + \Delta(G))$. We note that by the symmetry of the process, it suffices to show that the total amount of work over all calls of the form SMOOTHRHDIST $((i, F(i)), G, \epsilon)$ is $\mathcal{O}(\Delta(F) + \Delta(G))$. For clarity of notation, we will denote by ϵ_i the value of ϵ passed to SMOOTHRH when processing the pair $(i, F(i))$.

It is relatively easy to see that computational cost of the while loop at line 17 over all i is at most $\mathcal{O}(\Delta_F + \Delta_G)$ by the monotonicity of the right edge $\text{box}((i, F(i)); \epsilon_i)$. Specifically, as ϵ_i is a monotonically increasing function of i, so is $r_i = \lfloor (1 + \epsilon_i)i \rfloor$. But then, other than a constant amount of work associated with each iteration i, this implies that for every $1 \leq k \leq \Delta(G) + 1$, the value of $G(k)$ is retrieved at most once. Thus the total accesses to the ccdh of G are bounded by $\mathcal{O}(\Delta(F) + \Delta(G))$ as desired.

Now we consider the case where $\text{box}((i, F(i)); \epsilon_i)$ is above the smooth ccdh for G. We note that in this case $\ell_i = \lceil (1 - \epsilon_i)i \rceil$ is not a monotonic function of i, so we will have to apply different arguments. To that end, let $j = \lceil (1 - \epsilon_{i+1})i \rceil$, i.e., the value of j which terminates the while loop on line 7. Suppose by way of contradiction that $1 < j < i$. First we note that $\text{box}((j, G(j)); \epsilon_{j+1})$ and $\text{box}((j - 1, G(j - 1)); \epsilon_j)$ both contain the smooth ccdh for F and furthermore, the smooth ccdh of F must lie in the union of halfspaces

$$\mathcal{U} = \{(x, y) \colon y \geq F(i)\} \cup \{(x, y) \colon y \geq -F(i)(x - i) + F(i)\}.$$

Further, $j - 1 < (1 - \epsilon_{i+1})i$, we have that

$$(1 + \epsilon_j)(j - 1) \leq (1 + \epsilon_i)(j - 1) < (1 + \epsilon_i)(1 - \epsilon_{i+1})i < (1 + \epsilon_{i+1})(1 - \epsilon_{i+1}) < i.$$

Thus, in order for $\text{box}((j-1, G(j-1)); \epsilon_{j-1})$ to contain the smooth ccdh of F, we have that

$$F(i) \leq (1 + \epsilon_{j-1})G(j-1) \leq (1 + \epsilon_i)G(j-1).$$

In a similar manner, we have that

$$-F(i)\left((1 + \epsilon_i)j - i\right) + F(i) \leq (1 + \epsilon_i)G(j).$$

Now letting $\delta = (1 - \epsilon_{i+1})i - (j-1)$, we have by the definition of ϵ_{i+1} that

$$(1 - \epsilon_{i+1})F(i) = G(j-1) + \delta\left(G(j) - G(j-1)\right).$$

Combining these inequalities we get that

$$\begin{aligned}
(1 + \epsilon_i)\left(1 - \epsilon_{i+1}\right)F(i) &= (1 + \epsilon_i)\left(1 - \delta\right)G(j-1) + (1 + \epsilon_i)\,\delta G(j) \\
&\geq (1 - \delta)\,F(i) + \delta\left[-F(i)\left((1 + \epsilon_i)j - i\right) + F(i)\right] \\
&= \left[(1 - \delta) - \delta(1 + \epsilon_i)j + \delta i + \delta\right]F(i) \\
&= \left[1 + \delta i - \delta(1 + \epsilon_i)j\right]F(i).
\end{aligned}$$

As $F(i) > 0$ and $j = (1 - \epsilon_{i+1})i + (1 - \delta)$, this implies that

$$(1 + \epsilon_i)(1 - \epsilon_{i+1}) \geq 1 + \delta\left[i - (1 + \epsilon_i)(1 - \epsilon_{i+1})i + (1 - \delta)(1 + \epsilon_i)\right].$$

Since $\epsilon_i < \epsilon_{i+1}$, we have that $(1 + \epsilon_i)(1 - \epsilon_{i+1}) < 1$. But this yields that the left-hand side of the above inequality is strictly smaller than 1, while the right hand side is strictly larger than 1, a contradiction.

Thus, the while loop on line 7 can only terminate when $j = i$ or $j = 1$. The second case represents only a constant amount of work per iteration, contributing at most $\Delta(F)$ to the overall run time. For the first, we notice that this occurs only if $G(1) < \frac{1}{i}F(i)$. In this case, $\epsilon_{i+1} = 1 - \frac{G(1)}{F(i)}$ and in particular for all $k \geq i + 1$,

$$(1 - \epsilon_k)F(k) \leq (1 - \epsilon_{i+1})F(i) = G(1). \qquad \blacksquare$$

But this implies that line 7 can terminate with $j = 1$ precisely once when called as a subroutine of SMOOTHRH, and hence the total running time is $\Theta\left(\Delta(F) + \Delta(G)\right)$, as desired.

Lastly, we examine the time complexity of SMOOTHRH in practice. In Figure 1(b), we plot wall times for RH distance computation on 40 graph datasets collected from SNAP [23], yielding a total of $\binom{40}{2} = 780$ pairs of graphs. This experiment was run on a 2015 Macbook Pro laptop with a 2.8 GHz Intel Core i7 processor using a Python implementation of algorithm SMOOTHRH. Observe that the linear relationship apparent in Figure 1(b) is congruent with the time complexity stated in Theorem 3.2. In the next section, we present the RH distance values between these graphs.

4. Examples and the triangle inequality. In introducing both discrete [33] and smooth [26] RH distance, the authors examine RH distance on simple families of graphs to illustrate its behavior and compare it with other degree-based comparison metrics. In this section, we present examples of RH distance between real graph datasets and analyze RH distance between structured families of graphs.

Figure 2 depicts a heat map of RH distance values between pairs of graphs collected from SNAP [23]. The datasets are grouped by network type (as indicated by their SNAP label prefix), sorted within each group by maximum degree, and then sorted by group average maximum degree. The RH values range from 0.1 to 373, with mean 20 and median 8.49. The blue blocks on the diagonal identify network types with relatively small intragroup RH distance, such as road, peer-to-peer, and coauthorship networks. In contrast, citation and communication networks feature larger intragroup RH distance values. The red portions in the outer corners of the heat map indicate large RH distance between graphs with large differences in maximum degree. For instance, `roadNet-CA` and `com-Orkut` achieve the largest RH distance and have maximum degree 12 and 33,313, respectively. This pattern reflects that, while not solely controlled by maximum degree, RH distance is sensitive to differences in the tails of degree distributions.
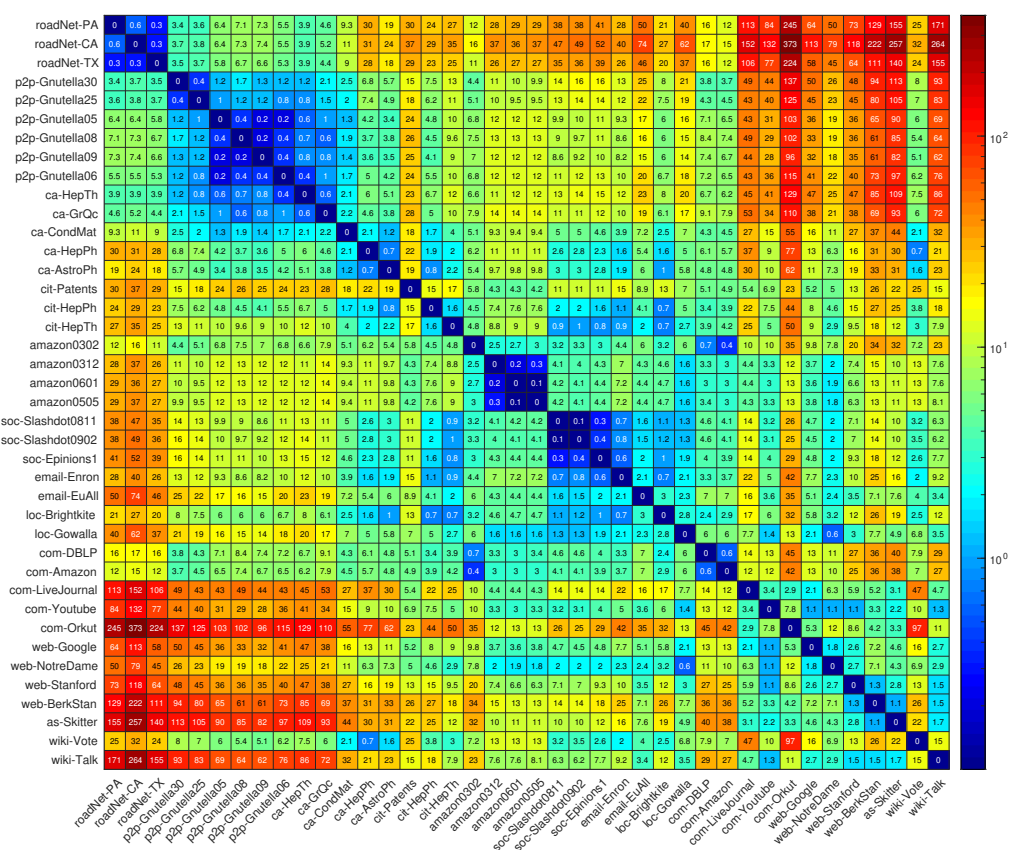

Figure 2. *Heat map of RH distance values between* 40 *graph datasets from SNAP* [23].

Next, we turn our attention to analyzing RH distance between structured families of graphs. To this end, we first establish a lemma useful for analytically determining RH distance between graphs whose degree sequence satisfies certain properties. Applying this lemma, we then find closed-form expressions for RH distance between complete graphs of different sizes and between the complete graph and cycle. Lastly, we conclude this section by analyzing these examples to show RH distance does not satisfy the triangle inequality, nor can the definition of RH distance be modified in certain ways to satisfy the triangle inequality.

Lemma 4.1. *Let F and G be graphs, and suppose that F has no vertices of degree k for $d \leq k \leq d'$. Let $1 \leq x, x' \leq \Delta(G) + 1$ and $\epsilon > 0$ such that*

$$|d - x| \leq \epsilon d, \qquad\qquad |d' + 1 - x'| \leq \epsilon(d' + 1),$$
$$|F(d) - G(x)| \leq \epsilon F(d), \qquad and \qquad |F(d' + 1) - G(x')| \leq \epsilon F(d' + 1).$$

Then for all $d < \delta < d' + 1$, G intersects $\mathrm{box}((\delta, F(\delta)); \epsilon)$.

Proof. We can choose x and x' such that $x \leq x'$, so assume they are chosen as such. Fix $d < \delta < d' + 1$ and note that since there are not vertices of degree between d and d', we have that there is some $y = F(d) = F(\delta) = F(d' + 1)$. By assumption we have that $\mathrm{box}((d, y); \epsilon)$ and $\mathrm{box}((d' + 1, y); \epsilon)$ both intersect the smooth ccdh of G. Noting that

$$(1 + \epsilon)F(\delta) \geq G(x) \geq G(x') \geq (1 - \epsilon)F(\delta),$$

G is monotonically decreasing, and $[x, x'] \cap [(1 - \epsilon)\delta, (1 + \epsilon)\delta] \neq \varnothing$, completes the proof. ∎

As a practical consequence of Lemma 4.1, it suffices to consider the end points of the plateaus present in the smooth ccdh of F when calculating $\overrightarrow{\mathcal{RH}}(F, G)$. This observation makes it a straightforward exercise to verify the following result.

Lemma 4.2. *Let $3 \leq n \leq m$; then $\mathcal{RH}(K_n, K_m) = \frac{m-n}{n}$.*

Proof. By Lemma 4.1 it suffices to consider the RH distance from $(1, n)$ and $(n - 1, n)$ to the ccdh of K_m and from $(1, m)$ and $(m - 1, m)$ to ccdh of K_n. We note that for any $\delta > 0$, the upper right corner of $\mathrm{box}\left((1, n); \frac{m-n}{n} - \delta\right)$ is $\left(\frac{m}{n} - \delta, m - \delta n\right)$. As $\frac{m}{n} - \delta < m - 1$ and $m - \delta n < m$, this implies that $\mathcal{RH}(K_n, K_m) \geq \frac{m-n}{n}$. To show that $\mathcal{RH}(K_n, K_m) \leq \frac{m-n}{n}$, we first note that $\left(1 - \frac{m-n}{n}\right) m = 2m - \frac{m^2}{n} \leq n$, and so $\mathrm{box}\left((x, m); \frac{m-n}{n}\right)$ contains a point on the ccdh of K_n as long as $(1 - \frac{m-n}{n})x \leq n - 1$. This is obvious for $x = 1$, and for $x = m - 1$ the inequality $(2n - m)(m - 1) \leq n(n - 1)$ can be rearranged in to $m - n \leq (m - n)^2$ which clearly holds. Now we note that $\mathrm{box}\left((x, n); \frac{m-n}{n}\right)$ contains (x, m) on the boundary, and thus $\mathcal{RH}(K_n, K_m) = \frac{m-n}{n}$. ∎

Along similar lines it is relatively straightforward but tedious to derive closed-form expressions of RH distance between different families of graphs, such as the complete graph and cycle.

Lemma 4.3. *Let* $3 \leq n, m$; *then*

$$\mathcal{RH}(K_n, C_m) = \begin{cases} \frac{n}{m} - 1, & 3 \leq m < \frac{n^2 - 3n + \sqrt{n^4 + 2n^3 - 11n^2}}{4n - 10}, \\ 1 - \frac{3m}{n + nm - m}, & \frac{n^2 - 3n + \sqrt{n^4 + 2n^3 - 11n^2}}{4n - 10} \leq m \leq n, \\ 1 - \frac{3m}{n + nm - m}, & 5 \leq n < m \leq \frac{n^2 - 3n + \sqrt{n^4 - 4n^3 + 7n^2}}{n - 1}, \\ \frac{m}{n} - 1, & \frac{n^2 - 3n + \sqrt{n^4 - 4n^3 + 7n^2}}{n - 1} < m \leq 2n, \\ \frac{3m}{n + m} - 1, & 2n < m. \end{cases}$$

In addition to illustrating the nuance of RH distance values even for simple examples like the complete graph and cycle, such closed-form expressions are also useful in revealing mathematical properties of RH distance. For instance, a cursory analysis of the closed-form expression in Lemma 4.2 reveals that, while the RH measure clearly satisfies the nonnegativity and symmetry axioms defining a distance metric, RH distance does not satisfy the triangle inequality: observe that $\mathcal{RH}(K_3, K_4) + \mathcal{RH}(K_4, K_5) = 7/12 < 2/3 = \mathcal{RH}(K_3, K_5)$. A natural, subsequent question is whether RH distance satisfies a *c-relaxed triangle inequality*, a triangle inequality variant that has been utilized in pattern matching [15], of the form

$$\mathcal{RH}(X, Y) + \mathcal{RH}(Y, Z) \geq c \cdot \mathcal{RH}(X, Z),$$

for some fixed $c \leq 1$. However, a closer analysis of the closed-form expression in Lemma 4.2 reveals that a c-relaxed triangle inequality is also impossible.

Claim 4.4 (relaxed triangle inequality impossible). *For any* $\epsilon > 0$, *there exist graphs* F, G, H *such that*

$$\frac{\mathcal{RH}(F, G) + \mathcal{RH}(G, H)}{\mathcal{RH}(F, H)} < \epsilon.$$

Proof. We take F, G, and H as the complete graph on x, y, and cx vertices, respectively, for some constant $c > 1$ where $x < y < cx$. Using the closed formula for the RH distance between complete graphs provided in Lemma 4.2, the ratio in question is

$$f(y) := \frac{1}{c - 1} \left(\frac{y}{x} + \frac{cx}{y} - 2 \right),$$

and choosing $y = \sqrt{c}x$ which minimizes f, we see that $f(\sqrt{c}x) \to 0$ as $c \to \infty$. ∎

Lastly, we consider the possibility that the definition of RH distance itself could be modified in some way so that it satisfies the triangle inequality. Recall that the RH distance between graphs G and F is the maximum of the two directional RH distances between G and F. Instead of taking the extreme values of directional RH distance, one approach may be to "regularize" RH distance by taking some weighted average of directional RH distances. However, we provide a simple counterexample below to show no positive linear combination of maximum or minimum directional RH distances defines a bona fide distance metric.

Claim 4.5. *There exists no positive linear combination of* $\min\{\overrightarrow{\mathcal{RH}}(F, G), \overrightarrow{\mathcal{RH}}(G, F)\}$ *and* $\max\{\overrightarrow{\mathcal{RH}}(F, G), \overrightarrow{\mathcal{RH}}(G, F)\}$ *which defines a distance metric on graphs.*
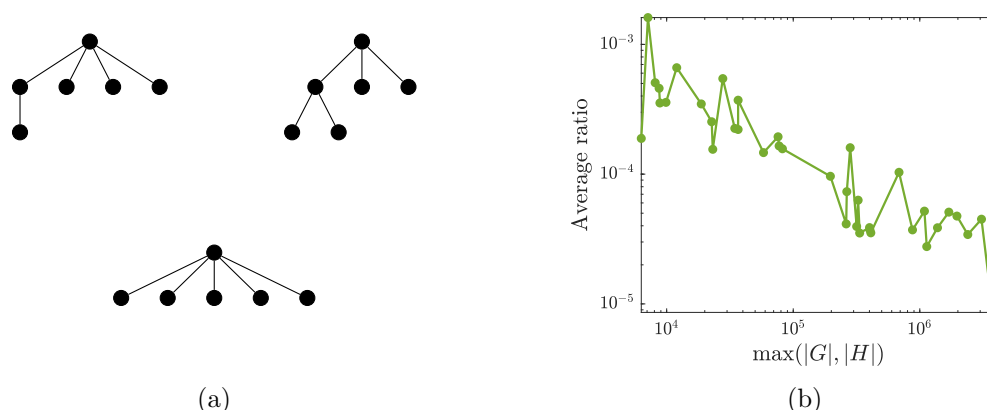
Figure 3. *Trees G (top left), H (top right), and F (bottom) (a). The ratio of $\mathcal{RH}(G, H)$ to the max possible RH distance, averaged over all pairs G, H in the SNAP data with $m = \max(|G|, |H|)$ (b).*

Proof. Consider the three trees on 6 vertices illustrated in Figure 3(a). A straightforward computation shows

$$\overrightarrow{\mathcal{RH}}(F, G) = \text{\textonehalf}/3 \text{ and } \overrightarrow{\mathcal{RH}}(G, F) = 1/12,$$
$$\overrightarrow{\mathcal{RH}}(F, H) = 1 \quad \text{and} \quad \overrightarrow{\mathcal{RH}}(H, F) = 1/2,$$
$$\overrightarrow{\mathcal{RH}}(H, G) = 1/5 \text{ and } \overrightarrow{\mathcal{RH}}(G, H) = 1/7$$

from which we observe that no positive linear combination of the maximum and minimum of the directional RH measures will always satisfy the triangle inequality since for any $a, b > 0$, we have $2(a + b/2) > a(\frac{1}{3} + 1 + \frac{1}{5}) + b(\frac{1}{12} + \frac{1}{2} + \frac{1}{7})$. ∎

In short, the notion of RH distance has seemingly intrinsic barriers to satisfying the triangle inequality. We note that RH distance's status as a bona fide distance metric may not solely be a question of theoretical interest since performing graph distance computations over objects embedded in a metric space can sometimes have practical benefits, such as faster graph algorithms [7]. Next, we consider other analytic properties of RH distance useful for understanding its behavior.

5. Range, density, and extremal properties of RH distance. In this section, we derive several analytic results on RH distance which characterize the range, density, and extremal behavior of RH distance values.

5.1. Range. A number of distance metrics, such as KS distance, are (by definition) bounded between 0 and 1. This normalization is a desirable property and facilitates interpretability when comparing distances between pairs of graphs of varying sizes. In the case of RH distance, while Simpson et al. [33] note that "RH distance can exceed 1," no upper bound is known. Below, we establish a sharp upper bound on RH distance.

Theorem 5.1. *Let $\mathcal{G}_k$ denote the set of all graphs on k vertices. If $n \leq m$, then*

$$\max_{\substack{F \in \mathcal{G}_n \\ G \in \mathcal{G}_m}} \mathcal{RH}(F,G) = \begin{cases} \frac{m}{2} - 1, & n = 2, \\ \frac{m^2}{2m+1} - 1, & n \geq 3, m \geq 4. \end{cases}$$

Proof. We first consider the case where $n = 2$ and $m \geq 3$. We note that K_2 is the unique graph on two vertices with no isolated vertices. Thus since $\{(1,i) \mid i \in [2,m]\} \subset \mathrm{box}\big((1,2); \frac{m}{2} - 1\big)$, we have that

$$\max_{\substack{F \in \mathcal{G}_2 \\ G \in \mathcal{G}_m}} \mathcal{RH}(F,G) > \frac{m}{2} - 1$$

if and only if there is some $G \in \mathcal{G}_m$ such that $\overrightarrow{\mathcal{RH}}(G,K_2) > \frac{m}{2} - 1$.

Now we note that $\mathrm{box}\big((x,G(x)); \frac{m}{2} - 1\big)$ contains the point $(1,2)$ for all $x \in Z_{\geq 1}$ when $m \neq 3$. Thus, we are left to consider when $m = 3$. In this case, $\overrightarrow{\mathcal{RH}}(P_3, K_2) = \overrightarrow{\mathcal{RH}}(K_3, K_2)$ $= \frac{1}{3}$. Thus, there is no $G \in \mathcal{G}_m$ such that $\overrightarrow{\mathcal{RH}}(G,K_2) > \frac{m}{2} - 1$. Finally, it is easy to verify that $\overrightarrow{\mathcal{RH}}(K_2, K_m) = \frac{m}{2} - 1$, which completes the case when $n = 2$.

For ease of notation, in the remainder of the proof we will define ϵ_m as $\frac{m^2}{2m+1} - 1$. Suppose now that F has exactly one vertex of degree greater than 1 and that $G = K_m$. Then

$$\mathrm{box}((2,1); \epsilon_m) = \left[2 - 2\epsilon_m, \frac{2m^2}{2m+1}\right] \times \left[1 - \epsilon_m, \frac{m^2}{2m+1}\right]$$

which intersects the ccdh for G precisely on the line between $(m-1, m)$ and $(m,0)$. Thus $\mathcal{RH}(F,G) \geq \epsilon_m$. Thus it suffices to show that for arbitrary choices of F and G, $\mathcal{RH}(F,G) \leq \epsilon_m$.

Now consider the case where $3 \leq n$ and $5 \leq m$. Since $\epsilon_m > 1$ we have that for any $(x,y) \in \mathbb{Z}^2$, $\mathrm{box}((x,y); \epsilon_m)$ contains the set $[0,m] \times \{y\}$ if $x > 2$ and the set $\{x\} \times [0,m]$ if $y > 2$. In particular, this implies that if $F \in \mathcal{G}_n$ and $G \in \mathcal{G}_m$, then $\mathcal{RH}(F,G) \leq \epsilon_m$ unless one of $F(2)$ or $G(2)$ is in $\{1,2\}$, as both $\mathrm{box}((1, G(1); \epsilon_m)$ and $\mathrm{box}((1, F(1)); \epsilon_m)$ contain the points $(1,n)$ and $(1,m)$ since $3 \leq n, m$. We now consider the intersection of the final line segment in a ccdh and the boxes around $(2,1)$ and $(2,2)$. To this end, let the endpoints of the final line segment be (Δ, c) and $(\Delta + 1, 0)$. Now we note that $[0, \frac{2m^2}{2m+1}] \times \{0\} \subset \mathrm{box}((2,1); \epsilon_m) \cap \mathrm{box}((2,2); \epsilon_m)$ and thus contains $(\Delta + 1, 0)$ unless $\Delta = m - 1$. It is then a simple algebraic exercise to show that the line segment $(m-1, c)$ to $(m,0)$ intersects $\mathrm{box}((2,1); \epsilon_m) \cap \mathrm{box}((2,2); \epsilon_m)$ for all $c \leq m$. As c is at most m by definition, this completes the this case.

The only remaining cases are when $m = 3, 4$. Consider first when $n = m = 3$. In this case, either the ccdhs are identical or one is $(3,1)$ and the other is $(3,3)$. Given that the only integer point these differ on is $d = 2$, it suffices to consider

$$\mathrm{box}((2,1); \epsilon_3) = \mathrm{box}\left((2,1); \frac{2}{7}\right) = [{}^{10}\!/_7, {}^{18}\!/_7] \times [{}^5\!/_7, {}^9\!/_7].$$

Noting that the point $({}^{18}\!/_7, {}^9\!/_7)$ is on the line segment between $(2,3)$ and $(3,0)$ completes this case.

Table 1
RH boxes around points relevant to maximum $\mathcal{RH}(F,G)$.

(x,y)	box$((x,y);\epsilon_4)$	Contains $(1,3)$	Contains $(1,4)$	Contains $[1,3] \times \{1\}$
$(1,3)$	$[2/9, 16/9] \times [6/9, 48/9]$	X	X	X
$(1,4)$	$[2/9, 16/9] \times [6/9, 64/9]$	X	X	X
$(2,1)$	$[4/9, 32/9] \times [2/9, 16/9]$			X
$(2,2)$	$[4/9, 32/9] \times [4/9, 32/9]$	X		X
$(2,3)$	$[4/9, 32/9] \times [6/9, 48/9]$	X	X	X
$(2,4)$	$[4/9, 32/9] \times [8/9, 64/9]$	X	X	X
$(3,1)$	$[6/9, 48/9] \times [2/9, 16/9]$			X
$(3,2)$	$[6/9, 48/9] \times [4/9, 32/9]$	X		X
$(3,4)$	$[6/9, 48/9] \times [8/9, 64/9]$	X	X	X

To consider the case where $m = 4$ and $n = 3, 4$, we first fix some $F \in \mathcal{G}_3 \cup \mathcal{G}_4$ and $G \in \mathcal{G}_4$ and suppose that $\mathcal{RH}(F,G) > \epsilon_4$. Noting that the only valid ccdhs on 4 vertices are $\{(4), (4,2), (4,4), (4,1,1), (4,3,1), (4,4,2), (4,4,4)\}$, it suffices to examine the boxes around a set of 9 points given in Table 1.

From Table 1 it is clear that we can restrict ourselves to the case where one of F or G passes through one of $(2,1), (2,2), (3,1)$, or $(3,2)$. Furthermore, since these boxes all contain the line segment $[1,3] \times \{1\}$ we have that the other graph must have at least two vertices of degree 3. Thus we may assume without loss of generality that the ccdh for F passes through one $S = \{(2,1), (2,2), (3,1), (3,2)\}$ and the ccdh for G is either $(4,4,2)$ or $(4,4,4)$. In particular, the ccdh for G at $\frac{32}{9}$ either passes through $8/9$ or $16/9$, respectively. As these points are in box$(s;\epsilon_4)$ for all $s \in S$, this completes the proof. ∎

We note the upper bound given by Theorem 5.1 allows one to define an RH *similarity* measure, in the spirit of cosine similarity, with values always in the interval $[0,1]$. We analyze how such a similarity measure might transform the RH distance values of the SNAP data by comparing each RH value with the maximum possible RH distance. That is, for each pair G, H of SNAP graphs, we consider the ratio

$$\frac{\mathcal{RH}(G,H)}{\frac{m^2}{2m+1} - 1},$$

where $m = \max(|G|, |H|)$. In Figure 3(b), we plot the mean of this ratio for each value of m. We observe RH distance values constitute a decreasing proportion of the maximum possible RH distance. Accordingly, in practice, normalizing RH distance values in this way will likely yield values that are vanishingly small in the size of the larger graph being compared. Hence, practitioners may choose to further transform such RH similarity values or simply opt to use (unnormalized) smooth RH distance instead.

5.2. Density. When quantifying the similarity between discrete objects like graphs, it is important to understand the breadth of possible values that can be assumed. In the case of the RH distance between graphs, such questions are more meaningful when considered asymptotically, since it is trivially true that for any fixed n, there are a finite number of n-vertex graphs and hence a limited number of possible RH distance values. Below, we show that RH distance is dense in the real numbers, in the following sense.

Theorem 5.2. *Let $c \in (0, \frac{1}{2}]$. There exists a sequence of pairs of graphs $\{(F_{n_i}, G_{n_i})\}_i$, each on n_i vertices, such that*

$$\lim_{i \to \infty} \frac{\mathcal{RH}(F_{n_i}, G_{n_i})}{n_i} = c.$$

Proof. First we note that $\frac{1}{c}$ can be expressed as $k + \delta$ where k is a positive integer that is at least 2 and $\delta \in [0, 1)$. Now let $\{n_i\}$ be the sequence of odd positive integers such that $2k + 4 \leq n_1$ and $\lceil \frac{\delta n_i}{k+\delta} \rceil = \lceil c\delta n_i \rceil$ is even. We note that this sequence is infinite and has a bounded gaps between successive entries. In particular, for every (sufficiently large) positive integer t there exists an odd n such that $\lceil \frac{\delta n}{k+\delta} \rceil = 2t$. To see this, we note that it suffices to show that there is an odd integer such that $(2t - 1) < \frac{\delta}{k+\delta} n \leq 2t$. Rearranging this expression, we get

$$2t \left(1 + \frac{k}{\delta} \right) - \left(1 + \frac{k}{\delta} \right) < n \leq 2t \left(1 + \frac{k}{\delta} \right).$$

As the gap between the upper and lower bounds on n is $1 + \frac{k}{\delta} > 3$, there is at least one odd integer for every choice of t and the gap between successive such odd integers is bounded by a function of k and δ.

We now define G_{n_i} as a r_i-regular graph on n_i vertices, where $r_i = \lfloor \frac{k}{k+\delta} n_i \rfloor - 1$. Note that $n_i - r_i = n_i - \lfloor \frac{k}{k+\delta} n_i \rfloor + 1 = \lceil \frac{\delta}{k+\delta} n_i \rceil + 1$, which is odd. Thus n_i and r_i have opposite parities, and so $r_i n_i$ is even and such a graph exists. We form F_{n_i} by creating a $(k-2)$-regular graph on $n_i - 1$ vertices and adding a dominating vertex. Note that the ccdh of G_{n_i} is n_i repeated r_i times while the ccdh of F_{n_i} is n_i repeated $k - 1$ times followed by 1, $n_i - k$ times.

We will show that $\mathcal{RH}(F_{n_i}, G_{n_i}) = \tau$ where $\tau = \frac{(r_i - k + 1)n_i - 1}{kn_i + 1}$. We first note that

$$r_i \geq \lfloor \frac{k(2k+4)}{k+\delta} \rfloor - 1 \geq \left\lfloor \frac{2k^2 + 4k}{k+1} \right\rfloor - 1 = 2k,$$

and thus the ccdhs G_{n_i} and F_{n_i} coincide on the interval $[1, k-1]$. Thus by Lemma 4.1, in order to show that $\mathcal{RH}(F_{n_i}, G_{n_i}) \leq \tau$, it suffices to show that F_{n_i} intersects $\text{box}((r_i, n_i); \tau)$ and that G_{n_i} intersects both $\text{box}((k, 1); \tau)$ and $\text{box}((n_i - 1, 1); \tau)$. Now as $r_i - k + 1 \geq 2k - k + 1 = k + 1$, we have that $\tau > 1$. Thus, trivially, F_{n_i} intersects $\text{box}((r_i, n_i); \tau)$ and G_{n_i} intersects $\text{box}((n_i - 1, 1); \tau)$. Now consider $\text{box}((k, 1); \tau)$ and note that its upper right coordinate is

$$\left(\frac{(r_i + 1) n_i}{kn_i + 1} k, \frac{(r_i + 1) n_i}{kn_i + 1} \right)$$

Since

$$r_i < \frac{(r_i + 1) n_i}{kn_i + 1} k < r_i + 1$$

and

$$-n_i \left(\frac{(r_i + 1) n_i}{kn_i + 1} k - r_i \right) + n_i = -n_i \frac{kn_i - r_i}{kn_i + 1} + n_i = \frac{-kn_i^2 + r_i n_i + kn_i^2 + n_i}{kn_i + 1} = \frac{(r_i + 1) n_i}{kn_i + 1},$$

we have that $\text{box}((k, 1); \tau)$ touches G_{n_i} on the line between (r_i, n_i) and $(r_i + 1, 0)$. Further, since $\frac{(r_i+1)n_i}{kn_i+1} < n_i$, this implies that $\mathcal{RH}(F_{n_i}, G_{n_i}) = \tau$.

Finally, noting that

$$\lim_{i \to \infty} \frac{1}{n_i} \frac{(r_i - k + 1)n_i - 1}{kn_i + 1} = \lim_{i \to \infty} \frac{\left(\left\lfloor \frac{k}{k+\delta} n_i \right\rfloor - k \right) n_i - 1}{n_i(kn_i + 1)} = \frac{1}{k+\delta} = c$$ ∎

completes the proof.

5.3. Extremal edge deletion. In this section, we aim to uncover which minimal change in graph structure results in a maximal change in RH distance. On this topic, recall that in [26, 35] Matulef and Stolman give an example of two n-vertex graphs differing only by a single edge, yet with nonvanishing discrete RH distance and vanishing smooth RH distance, in order to to argue smooth RH distance fixes flaws in the discrete RH distance. Here, we investigate to what extent this property of (smooth) RH distance holds, not only for their particular example, but in general. Namely, we ask, over all possible pairs of n-vertex graphs that differ by an edge, what is the maximum (smooth) RH distance, and which pair of graphs achieves this maximum?

As we soon will show, an RH extremal pair of edge-differing graphs is given by the star graph[3] and a "perturbed" star graph containing an additional edge between pendant (i.e., degree 1) vertices. More generally, the extremal pair is given precisely by any such "star-degreed" graph, i.e., the following.

Definition 5.3. *An n-vertex graph with precisely one degree $k \geq 2$ vertex and $n-1$ pendant vertices is called star-degreed.*

Besides the star graph itself (on at least 3 vertices), disconnected graphs consisting of the disjoint union of a star on $k \geq 3$ vertices and $n-k/2$ many disjoint edges are also star-degreed. Below, we calculate the discrete and smooth RH distance between a star-degreed graph S and graph $S^* = S \cup e$ where e is any edge between pendant vertices in S.

Claim 5.4 (star-degreed vs. perturbed star-degreed). *For $n \geq 5$, let S be a star-degreed graph, and let $S^* = S \cup e$ where e is any edge between pendant vertices in S. Then*
  (i) $\overrightarrow{\mathcal{RH}}_d(S, S^*) = 1/2$ *and* $\overrightarrow{\mathcal{RH}}_d(S^*, S) = 2/3$;
  (ii) $\overrightarrow{\mathcal{RH}}(S, S^*) = 2/5$ *and* $\overrightarrow{\mathcal{RH}}(S^*, S) = \frac{2}{2n+1}$.

Proof. Letting F and G denote the ccdhs of S and S^*, respectively, it is clear that $F(2) = 1$, $G(2) = 3$, and $F(d) = G(d)$ for all other $d \in \mathbb{Z}_{\geq 1}$. We first prove (i). Fixing $d = 2$, we note that $\epsilon^* = \frac{1}{2}$ is the minimum ϵ such that the set of integers $\{\lceil(1-\epsilon)d\rceil, \dots, \lfloor(1+\epsilon)d\rfloor\} \neq \{2\}$. Then since $\{\lceil(1-\epsilon^*)d\rceil, \dots, \lfloor(1+\epsilon^*)d\rfloor\} = \{1, 2, 3\}$ and

$$\min_{d' \in \{1,2,3\}} \frac{|F(2) - G(d')|}{F(2)} = 0 < \frac{|F(2) - G(2)|}{F(2)} = 2,$$

we have that $\overrightarrow{\mathcal{RH}}_d(S, S^*) = 1/2$. Now, switching the roles of F and G, we observe $|G(2) - F(d')|/G(2) \geq \frac{2}{3}$ for all d' when $n \geq 5$, with equality holding if $d' = 2$; hence $\overrightarrow{\mathcal{RH}}_d(S^*, S) = 2/3$.

---

[3]Recall that a star graph on n vertices is the complete bipartite graph, $K_{1,n-1}$.

Next, we prove (ii). Since for $d' \in [1, 2]$, we have

$$\frac{|F(2) - G(d')|}{F(2)} \geq \frac{|F(2) - G(2)|}{F(2)},$$

it suffices to restrict attention to $d' \in [2, 3]$. Rearranging $\frac{|F(2) - G((1+\epsilon)2)|}{F(2)} \leq \epsilon$ yields $\epsilon \geq 2/5$; hence $\overrightarrow{\mathcal{RH}}(S, S^*) = 2/5$. Now, switching the roles of F and G, we observe $|G(2) - F(d')|/G(2) = 2/3$ for all $d' \geq 2$. For $d' \in [1, 2]$, rearranging $\frac{|G(2) - F((1-\epsilon)2)|}{G(2)} \leq \epsilon$ yields $\epsilon \geq \frac{2}{2n+1}$; hence $\overrightarrow{\mathcal{RH}}(S^*, S) = \frac{2}{2n+1}$. ∎

Below, we prove that the star-degreed and perturbed star-degreed graphs considered above form the extremal pair maximizing RH distance over all n-vertex graphs differing by an edge.

Theorem 5.5. *Given $G = (V, E)$ and $e \notin E$, let $G \cup e$ denote the graph on V with $E = E(G) \cup e$. Then*

$$\max_G \mathcal{RH}(G, G \cup e) = 2/5,$$

which is achieved by taking G to be a star-degreed graph and e as any edge between pendant vertices in G.

Proof. Let $G = (V, E)$ be a graph of order n and $e = xy \notin E$. Let F denote the ccdh of $G \cup e$. Then

$$F(d) = G(d) + \delta_{d,d(x)+1} + \delta_{d,d(y)+1},$$

where $\delta_{i,j}$ denotes the Kronecker delta function. Thus, we are left to consider when d equals either $d(x) + 1$ or $d(y) + 1$. Without loss of generality, let $d = d(x) + 1$. Next note that for any graph H, $H(1) = |V(H)|$ and $H(k + 1) = H(k) - n_k$, for $k \in \mathbb{Z}_{\geq 1}$, where n_k is the number of vertices in H with degree k. From this observation, it follows that $G(d-1) = G(d) + n_{d-1} \geq F(d)$. Thus, since $G(d-1) \geq F(d) > G(d) \geq F(d) - 2$ there exists d' such that $d - d' \leq \frac{4}{5} \leq \frac{2}{5}d$ and $G(d') = F(d) - \frac{2}{5}$, and so $\overrightarrow{\mathcal{RH}}(F, G) \leq \frac{2}{5}$. Now from the observation above, it also follows that $F(d+1) \leq G(d)$. Thus, since $F(d+1) \leq G(d) < F(d) \leq G(d) + 2$, there exists d' such that $d' - d \leq \frac{2}{5}d$ and $F(d') = G(d) + \frac{2}{5}$, and so $\overrightarrow{\mathcal{RH}}(G, F) \leq \frac{2}{5}$, as well. Therefore $\mathcal{RH}(G, F) \leq \frac{2}{5}$. Now suppose that $\mathcal{RH}(G, F) = \frac{2}{5}$. Then from our calculations above, we observe that $d = 2$ and $G(d) = 1$, so G must be star-degreed. Further, in this case when, for all d' such that $|d - d'| \leq \frac{2}{5}d$, $|G(d) - F(d')| \geq \frac{2}{5}G(d)$, and so

$$\max_G \mathcal{RH}(G, G \cup e) = 2/5,$$

with equality achieved precisely when G has no isolated vertices and one nonpendant vertex. ∎

This result shows that RH distance between graphs that differ by a single edge can be nonvanishing as the maximum of the degree tends towards infinity. However, the extremal family of graphs given by Theorem 5.5 (i.e., star-degreed graphs) is clearly highly structured. To investigate whether edge perturbation can produce unusually large RH distance in practice, we again turn to the SNAP data. For each graph, we compute the maximum RH distance over all possible single edge deletions and additions; that is,

$$\max_{e \in E(\overline{G})} \mathcal{RH}(G, G \cup e) \quad \text{and} \quad \max_{e \in E(G)} \mathcal{RH}(G, G \setminus e),$$
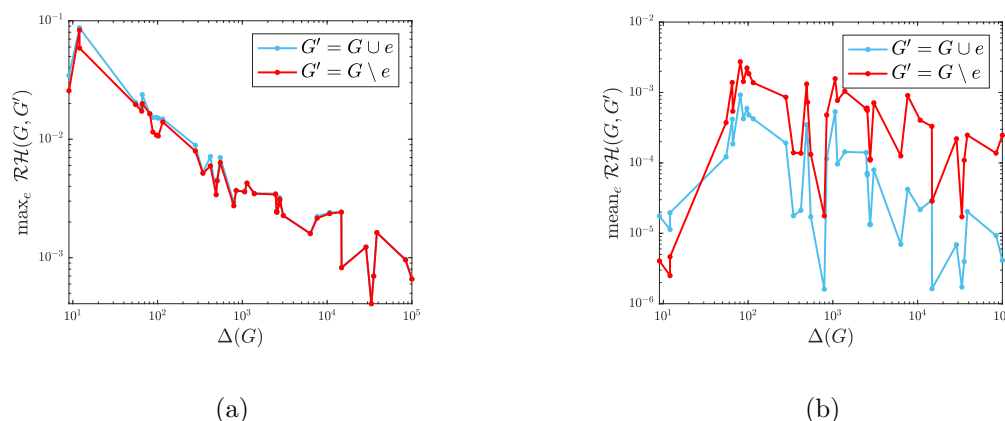
Figure 4. *The maximum* (a) *and average* (b) *RH distance produced by single edge additions and deletions on the SNAP graphs.*

where $\overline{G}$ denotes the complement of G. In addition to the maximum, we also computed the average RH distance over edge additions and deletions. Figure 4(a) shows the maximum RH distance between real graphs differing by an edge decreases in maximum degree of the graph. Figure 4(b) presents the average RH distance of single edge perturbations. While the same decreasing trend observed for the maximum is not apparent for the average, the results affirm that the expected RH distance produced by a random edge deletion or addition is small in practice. Both of these results suggest the behavior exhibited by star-degreed graphs is unlikely to be encountered in real data.

6. Conclusion and future work. We conducted an algorithmic and analytic study of graph RH distance, a newly proposed measure for quantifying similarity between degree distributions. Improving upon quadratic algorithms, we devised a linear-time algorithm for computing RH distance. Furthermore, we analyzed the basic properties of RH distance, answering a number of natural questions concerning RH distance's status as a bona fide metric, its range and density of possible values, and its extremal behavior with regard to single edge deletions. While RH distance has so far been mainly applied in the context of evaluating streaming and sampling algorithms [14, 26, 33, 35], it is by no means limited to these applications. Indeed, the work in [2] shows RH distance can be used as anomaly detection method in time-evolving graphs and is sometimes more sensitive to anomalies than the computationally expensive edit distance. Nonetheless, more work is needed to determine RH distance's utility in applications; in this regard, additional comparative studies of RH distance against more established metrics whose behavior is better understood would be valuable. Last but not certainly least, we have not explored RH distance in stochastic settings. What is the expected RH distance between two instantiations of a random graph model, or between the expected degree sequence and that of a realization of the random graph model? Such results would not only be of interest in themselves but provide a new way of quantifying the precision and accuracy of generative graph models in preserving degree distribution.

REFERENCES

[1] E. M. AIROLDI, X. BAI, AND K. M. CARLEY, *Network sampling and classification: An investigation of network model representations*, Decis. Support Syst., 51 (2011), pp. 506–518, https://doi.org/10.1016/j.dss.2011.02.014.

[2] S. G. AKSOY, K. E. NOWAK, E. PURVINE, AND S. J. YOUNG, *Relative Hausdorff Distance for Network Analysis*, e-print, 2019, https://arxiv.org/abs/1906.04936v1.

[3] S. G. AKSOY, E. PURVINE, E. COTILLA-SANCHEZ, AND M. HALAPPANAVAR, *A generative graph model for electrical infrastructure networks*, J. Complex Netw., 7 (2018), pp. 128–162, https://doi.org/10.1093/comnet/cny016.

[4] S. ALIAKBARY, J. HABIBI, AND A. MOVAGHAR, *Feature extraction from degree distribution for comparison and analysis of complex networks*, Computer J., 58 (2015), pp. 2079–2091, https://doi.org/10.1093/comjnl/bxv007.

[5] M. J. ATALLAH, *A linear time algorithm for the Hausdorff distance between convex polygons*, Inform. Process. Lett., 17 (1983), pp. 207–209, https://doi.org/10.1016/0020-0190(83)90042-x.

[6] A.-L. BARÁBASI AND R. ALBERT, *Emergence of scaling in random networks*, Science, 286 (1999), pp. 509–512.

[7] J. BENTO AND S. IOANNIDIS, *A family of tractable graph distances*, in Proceedings of the 2018 SIAM International Conference on Data Mining, SIAM, May 2018, pp. 333–341, https://doi.org/10.1137/1.9781611975321.38.

[8] V. D. BLONDEL, A. GAJARDO, M. HEYMANS, P. SENELLART, AND P. V. DOOREN, *A measure of similarity between graph vertices: Applications to synonym extraction and web searching*, SIAM Rev., 46 (2004), pp. 647–666, https://doi.org/10.1137/s0036144502415960.

[9] A. BRODER, R. KUMAR, F. MAGHOUL, P. RAGHAVAN, S. RAJAGOPALAN, R. STATA, A. TOMKINS, AND J. WIENER, *Graph structure in the web*, Computer Netw., 33 (2000), pp. 309–320, https://doi.org/10.1016/s1389-1286(00)00083-9.

[10] A. CLAUSET, C. R. SHALIZI, AND M. E. J. NEWMAN, *Power-law distributions in empirical data*, SIAM Rev., 51 (2009), pp. 661–703, https://doi.org/10.1137/070710111.

[11] R. COHEN, K. EREZ, D. BEN AVRAHAM, AND S. HAVLIN, *Resilience of the internet to random breakdowns*, Phys. Rev. Lett., 85 (2000), pp. 4626–4628, https://doi.org/10.1103/physrevlett.85.4626.

[12] L. DA F. COSTA, F. A. RODRIGUES, G. TRAVIESO, AND P. R. V. BOAS, *Characterization of complex networks: A survey of measurements*, Adv. Phys., 56 (2007), pp. 167–242, https://doi.org/10.1080/00018730601170527.

[13] J. C. DOYLE, D. L. ALDERSON, L. LI, S. LOW, M. ROUGHAN, S. SHALUNOV, R. TANAKA, AND W. WILLINGER, *The "robust yet fragile" nature of the Internet*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 14497–14502, https://doi.org/10.1073/pnas.0501426102.

[14] T. EDEN, S. JAIN, A. PINAR, D. RON, AND C. SESHADHRI, *Provable and practical approximations for the degree distribution using sublinear graph samples*, in Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18, ACM Press, 2018, https://doi.org/10.1145/3178876.3186111.

[15] R. FAGIN AND L. STOCKMEYER, *Relaxing the triangle inequality in pattern matching*, Int. J. Comput. Vis., 30 (1998), pp. 219–231, https://doi.org/10.1023/a:1008023416823.

[16] M. FALOUTSOS, P. FALOUTSOS, AND C. FALOUTSOS, *On power-law relationships of the internet topology*, in Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication - SIGCOMM '99, ACM Press, 1999, https://doi.org/10.1145/316188.316229.

[17] M.-L. FERNÁNDEZ AND G. VALIENTE, *A graph distance metric combining maximum common subgraph and minimum common supergraph*, Pattern Recogn. Lett., 22 (2001), pp. 753–758, https://doi.org/10.1016/s0167-8655(01)00017-4.

[18] F. Hausdorff, *Grundzuge der Mengenlehre*, American Mathematical Society, Providence, RI, 1914.

[19] J. Janssen, M. Hurshman, and N. Kalyaniwalla, *Model selection for social networks using graphlets*, Internet Math., 8 (2012), pp. 338–363, https://doi.org/10.1080/15427951.2012.671149.

[20] J. M. Kleinberg, *Authoritative sources in a hyperlinked environment*, J. ACM, 46 (1999), pp. 604–632, https://doi.org/10.1145/324133.324140.

[21] T. LaFond, J. Neville, and B. Gallagher, *Anomaly detection in networks with changing trends*, in Proceedings of the Outlier Detection and Description Under Data Diversity Workshop at the International Conference on Knowledge Discovery and Data Mining, 2014.

[22] S. H. Lee, P.-J. Kim, and H. Jeong, *Statistical properties of sampled networks*, Phys. Rev. E, 73 (2006), https://doi.org/10.1103/physreve.73.016102.

[23] J. Leskovec and A. Krevl, *SNAP Datasets: Stanford Large Network Dataset Collection*, June 2014, http://snap.stanford.edu/data.

[24] J. Leskovec and C. Faloutsos, *Sampling from large graphs*, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'06, ACM Press, 2006, https://doi.org/10.1145/1150402.1150479.

[25] C. L. Lin, *Hardness of approximating graph transformation problem*, in Algorithms and Computation, Springer, Berlin, 1994, pp. 74–82, https://doi.org/10.1007/3-540-58325-4_168.

[26] K. M. Matulef, *Final Report: Sampling-based Algorithms for Estimating Structure in Big Data*, technical report, Sandia National Laboratory, Livermore, February 2017, https://doi.org/10.2172/1367498.

[27] A. Montanari and A. Saberi, *The spread of innovations in social networks*, Proc. Natl. Acad. Sci. USA, 107 (2010), pp. 20196–20201, https://doi.org/10.1073/pnas.1004098107.

[28] S. Motallebi, S. Aliakbary, and J. Habibi, *Generative model selection using a scalable and size-independent complex network classifier*, Chaos, 23 (2013), 043127, https://doi.org/10.1063/1.4840235.

[29] M. E. J. Newman, *The structure and function of complex networks*, SIAM Rev., 45 (2003), pp. 167–256, https://doi.org/10.1137/s003614450342480.

[30] R. Pastor-Satorras and A. Vespignani, *Epidemic dynamics in finite size scale-free networks*, Phys. Rev. E, 65 (2002), https://doi.org/10.1103/physreve.65.035108.

[31] S. Ranshous, S. Shen, D. Koutra, S. Harenberg, C. Faloutsos, and N. F. Samatova, *Anomaly detection in dynamic networks: A survey*, Wiley Interdiscip. Rev. Comput. Stat., 7 (2015), pp. 223–247, https://doi.org/10.1002/wics.1347.

[32] A. Sanfeliu and K.-S. Fu, *A distance measure between attributed relational graphs for pattern recognition*, IEEE Trans. Systems Man Cybernet., SMC-13 (1983), pp. 353–362, https://doi.org/10.1109/tsmc.1983.6313167.

[33] O. Simpson, C. Seshadhri, and A. McGregor, *Catching the head, tail, and everything in between: A streaming algorithm for the degree distribution*, in Proceedings of the 2015 IEEE International Conference on Data Mining, November 2015, https://doi.org/10.1109/icdm.2015.47.

[34] A. Stolman, *HyperHeadTail*, https://github.com/astolman/HyperHeadTail/blob/master/scripts/rhcalculator.py, January 2017.

[35] A. Stolman and K. Matulef, *HyperHeadTail*, in Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 - ASONAM '17, ACM Press, 2017, https://doi.org/10.1145/3110025.3119395.

[36] M. P. H. Stumpf and C. Wiuf, *Sampling properties of random graphs: The degree distribution*, Phys. Rev. E, 72 (2005), https://doi.org/10.1103/physreve.72.036118.

[37] C. Yin, L. Yang, and J. Wang, *Botnet detection based on degree distributions of node using data mining scheme*, Int. J. Future Generation Commun. Netw., Networking, 6 (2013), pp. 81–90, https://doi.org/10.14257/ijfgcn.2013.6.6.09.