Aksoy *et al. Applied Network Science* (2019) 4:80
https://doi.org/10.1007/s41109-019-0198-0

Applied Network Science

RESEARCH Open Access

# Relative Hausdorff distance for network analysis

Sinan G. Aksoy[1*], Kathleen E. Nowak[1], Emilie Purvine[2] and Stephen J. Young[1]

*Correspondence:
sinan.aksoy@pnnl.gov
[1]Pacific Northwest National Laboratory, Richland, WA 99352, United States
Full list of author information is available at the end of the article

**Abstract**
Similarity measures are used extensively in machine learning and data science algorithms. The newly proposed graph Relative Hausdorff (RH) distance is a lightweight yet nuanced similarity measure for quantifying the closeness of two graphs. In this work we study the effectiveness of RH distance as a tool for detecting anomalies in time-evolving graph sequences. We apply RH to cyber data with given red team events, as well to synthetically generated sequences of graphs with planted attacks. In our experiments, the performance of RH distance is at times comparable, and sometimes superior, to graph edit distance in detecting anomalous phenomena. Our results suggest that in appropriate contexts, RH distance has advantages over more computationally intensive similarity measures.

**Keywords:** Relative Hausdorff distance, Graph similarity measure, Cyber anomaly detection, Temporal graphs

## Introduction

Similarity measures play a crucial role in many machine learning and data science algorithms such as image classification and segmentation, community detection, and recommender systems. A good deal of effort has gone into developing similarity measures for graphs, in particular, since they often provide a natural framework for representing unstructured data that accompanies many real-world applications. Some popular graph similarity measures currently used are graph edit distance (Sanfeliu and Fu 1983), iterative vertex-neighborhood identification (Blondel et al. 2004; Kleinberg 1999), and maximum common subgraph based distance (Fernández and Valiente 2001). However, as graph datasets grow larger and more complex, the need for tools that can *both* capture meaningful differences and scale well is becoming more critical. In this respect, a number of sophisticated yet costly graph similarity measures, such as those listed above, fall short.

The recently proposed graph Relative Hausdorff (RH) distance (Simpson et al. 2015) is a promising measure for quantifying similarity between graphs via their degree distributions. Inspired by the Hausdorff metric from topology (Hausdorff 1914), RH distance was devised to capture degree distribution closeness at all scales, and hence is well-suited for comparing the heavy-tailed degree distributions frequently exhibited by real-world graphs. Furthermore, as recent work has shown (Aksoy et al. 2018), RH distance is extremely lightweight, with time complexity linear in the maximum degrees of the graphs being compared. However, as this metric is relatively new, it has not yet been extensively vetted. In particular, current research has not addressed its potential as an anomaly detection method for time-evolving graphs.

In this work, we conduct a statistical and experimental study of RH distance in the context of dynamic graphs. While RH distance may be applied to arbitrary pairs of networks, we focus our attention on sequences of time-evolving networks arising from cyber-security applications. We begin by first applying RH distance to cyber-security logs recently released by Los Alamos National Laboratory, and investigate the extent to which it detects identified “red team" events. Then we follow up by studying RH distance in the more general and controlled context of random dynamic graph models. Here we generate sequences of correlated Chung-Lu random graphs using a simplified cyber-security model proposed by Hagberg et al. (2016), and test the extent to which RH distance detects several planted attack profiles. Throughout our analysis, we compare the performance of RH to that of more well-known graph similarity measures, such as edit distance and Kolmogorov-Smirnov distance of degree distributions. With this work, we better clarify the range of differences captured by RH, and also highlight its practical advantages and disadvantages over other methods.

## Preliminaries

### Graph similarity measures

Below we define the graph similarity measures we consider for anomaly detection in time-evolving graphs. We begin with the graph Relative Hausdorff distance, the primary focus of our study.

#### *Relative Hausdorff distance*

Originally introduced by Simpson et al. (2015), the Relative Hausdorff (RH) distance between graphs is a numerical measure of closeness between their complementary cumulative degree histograms (ccdh). More precisely, the (discrete) ccdh of a graph $G$ is defined as $(N(k))_{k=1}^{\infty}$, where $N(k)$ denotes the number of vertices of degree *at least* $k$. This is related to the commonly used *degree distribution*, which is defined as $(n(k))_{k=1}^{\infty}$ where $n(k)$ denotes the number of vertices with degree *exactly* $k$. Note that the ccdh and degree distribution are equivalent in the sense that each can be uniquely obtained from the other; nonetheless, for the purpose of this exposition, it is more convenient to work with the ccdh. Slightly abusing notation, we write $G(d)$ for a graph $G$ to mean the value of the ccdh of $G$ at $d$, and let $\Delta(G)$ denote the maximum degree of $G$. With these definitions in hand, the (discrete) RH distance between $F$ and $G$ is then defined as:

**Definition 1** ((Discrete) Relative Hausdorff distance (Simpson et al. 2015)) *Let $F, G$ be graphs. The discrete directional Relative Hausdorff distance from $F$ to $G$, denoted $\overrightarrow{\mathcal{RH}}(F, G)$, is the minimum $\epsilon$ such that*

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

*and $\mathcal{RH}(F, G) = \max\{\overrightarrow{\mathcal{RH}}(F, G), \overrightarrow{\mathcal{RH}}(G, F)\}$ is the discrete Relative Hausdorff distance between $F$ and $G$.*

In this paper, we compute RH distance using smoothed ccdhs, in which successive points are connected via line segments, as recommended by Matulef (2017); Stolman and

Matulef (2017). Specifically, the authors define the smooth ccdh of a graph $G$, $G(d) : \mathbb{R}_{\geq 1} \to \mathbb{R}_{\geq 0}$, as

$$G(d) = \begin{cases} \text{\# of vertices of degree at least} d, & d \in \mathbb{Z}_{\geq 1} \\ (d - \lfloor d \rfloor)G(\lfloor d \rfloor) + (\lceil d \rceil - d)G(\lceil d \rceil), & d \in \mathbb{R}_{\geq 1} \setminus \mathbb{Z}. \end{cases}$$

In this case, the RH distance is defined much the same as before, except that the ccdh is piecewise linear. An illustration of the RH equivalent of an $\epsilon$-ball at points on a smooth ccdh is given in Fig. 1 and the precise definition of smooth RH distance is given below. Henceforth, we focus exclusively on smooth RH distance so we will drop the qualifier.

**Definition 2** ((Smooth) Relative Hausdorff distance (Matulef 2017; Stolman and Matulef 2017)) *Let $F, G$ be graphs. The smooth directional Relative Hausdorff distance from $F$ to $G$, denoted $\overrightarrow{\mathcal{RH}}(F, G)$, is the minimum $\epsilon$ such that*

$$\forall d \in \{1, \ldots, \Delta(F)\}, \exists d' \in \mathbb{R}_{\geq 1} \textit{ such that} |d - d'| \leq \epsilon d \textit{ and} |F(d) - G(d')| \leq \epsilon F(d),$$

*and $\mathcal{RH}(F, G) = \max\{\overrightarrow{\mathcal{RH}}(F, G), \overrightarrow{\mathcal{RH}}(G, F)\}$ is the smooth Relative Hausdorff distance between $F$ and $G$.*

By definition, $\mathcal{RH}(F, G) = \epsilon$ means that for every degree $k$ in the graph $F$, $F(k)$ is within $\epsilon$-fractional error of $G(k')$ for some $k'$ within $\epsilon$-fractional error of $k$. Hence, 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 $\epsilon$-close to $G$ (and vice versa).

While RH distance was inspired by the Hausdorff distance metric (Hausdorff 1914) after which it is named, the concept underlying RH distance between graphs differs from

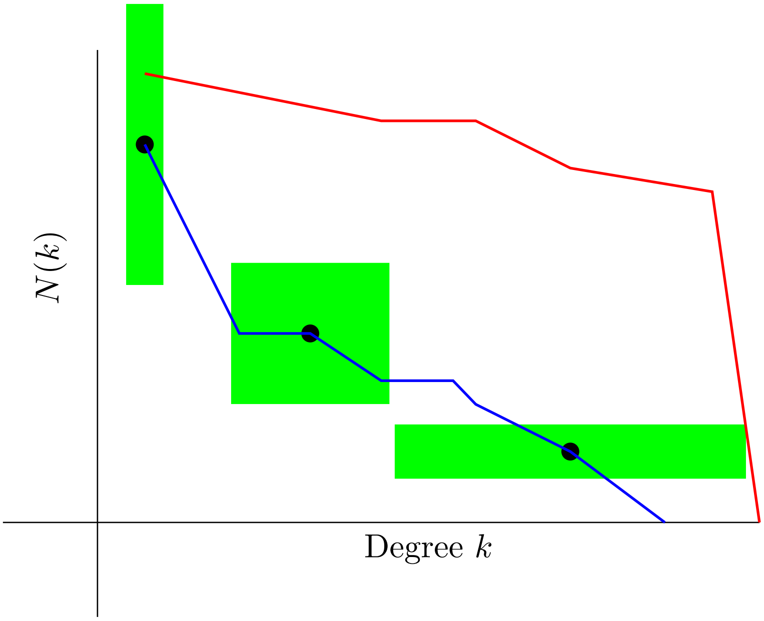


**Fig. 1** An illustration of the RH equivalent of an $\epsilon$-ball (in green) on one graph ccdh (in blue) compared with another (in red) at different points

Hausdorff distance in several important regards. Recall the directional Hausdorff distance from non-empty subset $X$ to $Y$ of a metric space $(M, d)$ is $\sup_{x \in X} \inf_{y \in Y} d(x, y)$, or equivalently, $\sup_{x \in X} \inf\{\varepsilon > 0 : B(x; \varepsilon) \cap Y \neq \varnothing\}$, where $B(x; \varepsilon)$ denotes the closed $\epsilon$-ball centered at $x$. Crucially, RH distance replaces $B(x; \varepsilon)$ with balls that are non-uniform in $X$, as illustrated in Fig. 1. This *relative* notion of ball, while no longer a true metric ball, is more appropriate for analyzing differences in highly-skewed degree distributions exhibited by complex networks. However, as discussed and analyzed in Aksoy et al. (2018), this also decouples the "distance" from the underlying topology, yielding a function $\mathcal{RH}(F, G)$ that does not satisfy the triangle inequality and hence is best viewed as a similarity measure rather than bona-fide distance metric. However, to match existing literature we will still use the term "RH distance." Lastly, it is worth noting $\mathcal{RH}(F, G)$ is extremely lightweight, and can be computed with run time $\mathcal{O}(\Delta(F) + \Delta(G))$; for a linear-time algorithm and more on theoretical properties of RH distance, the reader is referred to Aksoy et al. (2018).

### *Other measures*

While RH distance will be the focus of the present work, we also consider several other graph similarity measures in order to provide relevant context for its performance. First, we consider another comparable lightweight, ccdh-based measure called Kolmogorov-Smirnov (KS) distance. KS distance is a widely-used statistical measure of similarity between distributions, and serves as the test statistic for the two-sample KS hypothesis test (Gibbons and Chakraborti 2011; Young 1977). In what follows, we will not only compute KS distance directly between graph degree distributions, but also between distributions of graph similarity values, such as RH values. To avoid confusion, below we define both KS distance as well as the two-sample KS hypothesis test (for which KS distance is a test statistic) for general empirical distributions.

**Definition 3** (KS distance and two-sample KS test (Gibbons and Chakraborti 2011)) *Let F and G be empirical cumulative distribution functions formed from n and m samples, respectively. The Kolmogorov–Smirnov distance is*

$$\mathcal{KS}(F, G) = \max_x |F(x) - G(x)|.$$

*For the asymptotic null distribution, the null hypothesis that F and G are samples of two identical probability density functions is rejected at the α confidence level if*

$$\mathcal{KS}(F, G) > c(\alpha)\sqrt{\frac{n+m}{nm}},$$

*where* $c(\alpha)$ *is given asymptotically by* $\sqrt{-\frac{1}{2}\log\alpha}$.

See ((Gibbons and Chakraborti 2011), Ch. 6) and the references contained therein for a detailed overview of the two-sample KS test and derivation of the asymptotic expressions above. For small sample sizes, tables of critical values may be used in place of asymptotic estimates (e.g. for $n, m \leq 25$, see (Siegel and N.J.C. 1988), Table $L_I - L_{II}$). The reader is referred to Marsaglia et al. (2003); Simard and L'Ecuyer (2011) for further discussion on exact vs. approximate methods for computing the Kolmogorov-Smirnov distribution.

For clarity, we emphasize that smaller values of KS distance indicate greater *similarity* between distributions, whereas smaller $p$-values permit one to reject the null hypothesis of identical underlying distributions at a higher confidence level, thereby presenting stronger evidence the empirical distributions were drawn from *different* underlying distributions. For the special case that $F$ and $G$ are the ccdhs of two graphs on $n$ and $m$ vertices, respectively, KS distance is given by $\mathcal{KS}(F,G) = \max_{x \in \mathbb{N}} |\widetilde{F}(x) - \widetilde{G}(x)|$, where $\widetilde{F} = \frac{1}{n} \cdot F$ and $\widetilde{G} = \frac{1}{m} \cdot G$ . As argued in Matulef (2017); Stolman and Matulef (2017), KS distance between graph ccdhs can sometimes be large in graphs that are intuitively similar; furthermore, KS distance may also be insensitive to certain important differences between graph ccdhs, particularly in the tails of ccdhs (which correspond to high-degree vertices).

On the other end of the computational spectrum, we also consider *graph edit distance* (GED). Arguably one of the most well-known graph similarity measures, GED has been widely used throughout machine learning, particularly in computer vision and pattern recognition contexts (Gao et al. 2009). An *edit operation* on a graph consists of either an *insertion*, *deletion*, or *substitution* of a single vertex or edge.[1] An *edit path of length* $k$ between $F$ and $G$ is a sequence of edit operations $\mathcal{P} = (e_1, \ldots, e_k)$ that takes $F$ to a graph that is isomorphic to $G$. Edit distance is the total weight of the minimum-cost edit path, i.e.

**Definition 4** (Graph edit distance) *Let $F, G$ be graphs. The graph edit distance between $F$ and $G$, denoted $GED(F,G)$ is given by*

$$GED(F,G) = \min_{\mathcal{P} \in \Upsilon(F,G)} \sum_{e_i \in \mathcal{P}} c(e_i),$$

*where $\Upsilon(F,G)$ denotes the set of possible edit paths from $F$ to $G$, and $c(e_i) \geq 0$ denotes a cost-function measuring the weight of edit operation $e_i$.*

In what follows, we simply take $c(e_i) = 1$ for any edit operation, in which case $\text{GED}(F,G)$ is the minimum number of edit operations needed to transform $F$ to $G$.

### Related literature

While in this work we focus on the graph Relative Hausdorff distance, we note that a wide variety of graph similarity measures have been utilized for anomaly detection in time-evolving graphs. In Ishibashi et al. (2010), the authors propose detecting anomalies in communication network traffic data by measuring cosine similarity between the principal eigenvectors of graph adjacency matrices. In Akoglu and Faloutsos (2010), also take an eigenvector-based approach for measuring graph anomalousness. Matrix-analytic graph similarity measures have also been based on eigenvalue residuals (Giuseppe et al. 2011), non-negative matrix factorization (Tong and Lin 2011), and tensor decompositions (Sapienza et al. 2015). Other popular approaches for graph-based anomaly detection are via *distance metrics*, such as those based on edit distance, maximum common subgraph distance, or mean vertex eccentricity (Gaston et al. 2006), or take a community-detection approach towards identifying anomalies by tracking changes between clusters of well-connected vertices (Aggarwal et al. 2011; Wang and Paschalidis 2017). For a broader

[1] While others sometimes including *merging* and *splitting* edit operations, we restrict our attention to edit distance based on the three aforementioned operations.

survey of graph-based anomaly detection techniques see (Akoglu et al. 2014; Ranshous et al. 2015; Sensarma and Sarma 2015) and the references contained therein. Lastly, we note that applications of graph similarity functions extend far beyond anomaly detection. Graph similarity functions are also ubiquitous in inexact graph matching and graph classification problems. For instance, graph edit distance is a key tool for error-tolerant pattern recognition and computer vision techniques (Gao et al. 2009). While in this work we explore Relative Hausdorff distance through the lens of anomaly detection, we note its application as a graph similarity measure in other contexts such as these remains unexplored.

As our focus in this paper will be on *cyber* anomaly detection[2] in network flow data, we also mention some of the existing graph-based methods specifically for cyber anomaly detection. In this domain, some researchers focus on detecting anomalies edge-by-edge or target specific types of behavior, e.g., (François et al. 2011; Noble and Adams 2018), while others look at the graph more globally or structurally and are agnostic to the type of anomalous behavior being detected, e.g., (Chen et al. 2016). In Noble and Adams (2018) describe a real-time unsupervised framework for detecting anomalies in network data. They consider "edge activity" as the sequence of flows on a single edge and compute correlations between event inter-arrival times and other edge data (e.g., byte count or protocol). Statistically significant changes in those correlations are flagged as anomalies. Groups of adjacent anomalies can be combined to form larger anomalies perhaps indicating coordinated behavior. The authors of François et al. (2011) use PageRank to perform linkage analysis followed by clustering techniques to identify groups of IPs with similar behavior. These groups are then compared with known bot behavior to detect botnets within the network. In the category of more structural and behavior-agnostic algorithms (Chen et al. 2016) introduces multi-centrality graph PCA and multi-centrality graph dictionary learning which use structural properties of a graph, e.g., walk statistics and centrality measures, to learn normal structure and thus detect abnormal structure. This method is not tailored to the cyber use case, but the authors use network flow as one of their examples. Our work is similarly not targeted towards a specific cyber use case and is focused on detecting structural perturbations rather than clustering behavioral patterns.

Finally, we note others have measured network similarity using Hausdorff distance and the related Gromov-Hausdorff distance (Edwards 1975; Gromov 1981) on metric spaces. Banič and Taranenko (2015) define the Hausdorff distance between two simple, connected graphs based on the lattice of all subgraphs of the graphs in question. In Lee et al. (2011), a quantity inspired by Gromov-Hausdorff distance is applied to analyze brain networks. This quantity relies on the embedding of the network in a geometric structure; in general, Gromov-Hausdorff distance is defined over all isometric embeddings of a metric space. Recent work in Choi (2019) proposes fundamental definitions toward a theory of Gromov-Hausdorff distances between graphs, and includes exact calculations for a few simple classes of graphs. The wider applicability of this notion of Gromov-Hausdorff distance to real graph data is likely to face significant theoretical obstacles, as (Nowak et al.) shows that only a few special classes of graphs can be isometrically embedded in the class of metric spaces arising from a Hilbert space. Furthermore, both exact computation as well as approximation of Gromov-Hausdorff distance present computational challenges

[2] Note that we do not consider signature-based methods like those employed in intrusion detection/prevention systems (e.g., Snort) to be anomaly detection methods. Instead these are rule-based behavior identification tools.

(Agarwal et al. 2018). In contrast, we emphasize RH distance is applicable to *any* pair of graphs, admits linear-time computation, and (rather than requiring an associated graph embedding) is defined on graphs solely as abstract combinatorial objects.

## Los Alamos National Laboratory (LANL) cybersecurity data

To begin our study of RH distance as an anomaly detection method for dynamic graphs, we will first consider a dataset recently released by LANL with known red team events from their internal corporate computer network (Kent 2015a; 2015b). The dataset represents 58 consecutive days of de-identified event data collected from four sources, namely:

- Windows-based authentication events from both individual computers and centralized active directory domain controller servers,
- Process start and stop events from individual Windows computers,
- Domain Name Service (DNS) lookups collected by internal DNS servers, and
- Network flow data collected at several key router locations.

In total, the data set is approximately 87.4 gigabytes, spread across the four modalities, including 1,648,275,307 events coming from 12,425 users, 17,684 computers, and 62,974 processes. Ground truth for the red team events is given as a set of authentication events that are known red team compromise events. In this section, we will demonstrate that Relative Hausdorff distance is effectively able to identify anomalous behavior around the red team events in the LANL data.

### Data source

As stated above, LANL captured network evolution in four different modalities, namely authentication, process, network flow, and DNS events. In order to apply the RH distance, we first must convert these network event files into a time series of graphs. To do so, we consider 60 sec moving windows that advance 20 sec at a time. For each window we use the events in that window to construct a graph. For the 58 consecutive days, this yields a time sequence of 250,560 graphs for each modality. Further details for constructing each type of graph are given below.

- **Authentication Graphs.** The authentication data is a record of authentication events collected from individual Windows-based desktop computers, servers, and Active Directory servers. Each line of the data file reports a separate authentication event in the form
  ```
  time, sourceUser@domain, destUser@domain, source computer,
  dest computer,
  auth type, logon type, auth orientation, pass/fail.
  ```
  For a given window, we construct an unweighted graph with edges {sourceUser, destUser} for each user pair present in the logs within the window.
- **Authentication Failure Graphs.** These are constructed in the same manner as the Authentication Graphs, except we restrict the edge set to those corresponding to failed authentications only.
- **Process Graphs.** The process data is a record of process start and stop events collected from individual Windows-based desktop computers and servers. Each line of the data file reports a separate process start/stop in the form

`time, user@domain, computer, process name, start/end.`
For a given window, we construct an unweighted graph with edges {computer, process name} for each computer-process pair present in the logs within the window.

- **DNS Graphs.** The DNS data is a record of DNS lookup events collected from the central DNS servers within the network. Each line of the data file reports a separate lookup event in the form
  `time, source computer, computer resolved,`
  representing a DNS lookup at the given time by the source computer for the resolved computer. For a given window, we construct an unweighted graph with edges {source computer, computer resolved} for each source-resolved computer pair present in the logs within the window.
- **Flow Graphs.** The flow data is a record of the network flow events collected from central routers within the network. Each line of the data file reports a separate network flow event in the form `time, duration, source computer, source port, dest computer, dest port, protocol, packet count, byte count.`
  For a given window, we construct an unweighted graph with edges {source computer, dest computer} for each source-destination computer pair that communicate during that time.

### Limitations

While working with real-world data often presents challenges, testing graph-based anomaly detection methods on the LANL dataset is particularly difficult for several reasons. First and foremost, the data only provides red team authentication attempt time stamps and does not specify the nature, extent or duration of the red team events. This makes it difficult to segregate benign from anomalous time periods. Additionally, without knowing the specific red team actions, it is difficult to determine which (if any) of the aforementioned modalities a red team signature may appear in. Finally, it is worth noting the data exhibited large periods of time in which no events occurred that did not correspond to regular lulls such as weekends and nighttime. In particular, the flow data has records from only the first 37 of the 58 days. To address some of these limitations, in "Simulated evolving networks" section we extend our analyses to a generalized dynamic network model (Hagberg et al. 2016) proposed by LANL scientists Hagberg, Mishra, and Lemons. While no synthetic model is a perfect substitute for real data, this model's conception and design was directly informed by direct access to the LANL cyber data (Kent 2014; 2016) and provides a framework under which we may draw more certain and rigorous conclusions regarding the behavior of RH distance. First, we present our analysis of the real LANL data.

### Experiment and results

As a first-pass approach towards studying the sensitivity of RH distances to red team events in the LANL dataset, we test whether the distribution of pairwise RH distance values before a red team event differs significantly from the post red team event distribution. In this way, we assess whether there is statistical evidence to support that red team events demarcate “change-points” in RH distance distribution. To that end, for each red team event at time $r$, we associate a time window $w$ of length $\ell$ centered at $r$, which we denote $w_\ell(r)$. Each such window can be naturally partitioned into a “before” period (i.e. the time interval $(r - \ell/2, r)$) and “after” period, $(r, r + \ell/2)$. To avoid overlapping windows and ensure the “before" periods are in fact devoid of red team events, we restrict attentions to windows in which no red team event occurs in the time interval $(r - \ell/2, r)$. Put equivalently, we consider the set

$$W_\ell = \{w_\ell(r) : \text{red team event occurs at } r, \text{ no red team events occur in } (r - \ell/2, r)\}.$$

We note that it is possible for the after period of a window in $W_\ell$ to contain additional red team events. For each window in $W_\ell$, we compute the RH distances between pairs of graphs separated by $\delta$ seconds in the before period, as well as such pairs belonging to the after period. We then aggregate the RH distances over all before periods and all after periods. More precisely, if $G_0, G_1, \ldots$ denotes the time-ordered sequence of graphs for a particular mode in the LANL data, we compute the aggregate before and after distributions as

$$D_b = \{\mathcal{RH}(G_t, G_{t+\delta}) : t, t+\delta \in (r - \ell/2, r) \text{ and} w_\ell(r) \in W_\ell\},$$
$$D_a = \{\mathcal{RH}(G_t, G_{t+\delta}) : t, t+\delta \in (r, r + \ell/2) \text{ and} w_\ell(r) \in W_\ell\},$$

respectively. Recalling that we processed the LANL graph sequence for each modality by generating graphs for windows shifted by 20 sec, we may choose the parameter $\delta$ controlling the granularity of pairwise RH measurements to be as small as 20 sec and and as large as $\ell/2 - 20$ sec. Finally, we assess whether these aggregated before and after RH distance distributions differ significantly by conducting a two-sample Kolmogorov-Smirnov test. Table 1 presents the resulting $p$-values for $\delta = 20, 40, 60, 120, 240$ sec, under window lengths $\ell = 30, 60, 120$ min, for each LANL modality.

The $p$-values in Table 1 suggest that whether the aggregated distribution of RH values before red team events differs significantly from the post red team events depends crucially on the cyber modality, window length and granularity parameter $\delta$. In the case of a 30 min window, almost none of the parameter settings for any modality result in statistical significance, while for a 2-h window, a majority of parameter settings are significant at a level of 0.05. In this case, the before and after RH distance distributions over longer time windows surrounding red team events more frequently show significant differences, which is perhaps unsurprising. On the other hand, the changes in significance

**Table 1** The $p$-values of the two-sample KS test comparing RH distance distributions of aggregated before and after periods of time windows centered at red team events in the LANL data

| | Window: 30 min | | | | | Window: 60 min | | | | | Window: 120 min | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mode/Shift | 20s | 40s | 60s | 120s | 240s | 20s | 40s | 60s | 120s | 240s | 20s | 40s | 60s | 120s | 240s |
| AuthFail | 0.20 | 0.49 | 0.70 | 0.97 | 0.57 | 0.01 | 0.39 | 0.10 | 0.76 | 0.02 | 0.02 | 0.01 | 0.02 | 0.01 | 0.00 |
| Auth | 0.36 | 0.72 | 0.17 | 0.38 | 0.12 | 0.33 | 0.03 | 0.01 | 0.63 | 0.06 | 0.35 | 0.00 | 0.00 | 0.36 | 0.07 |
| Flow | 0.61 | 0.30 | 0.31 | 0.76 | 0.91 | 0.75 | 0.49 | 0.17 | 0.31 | 0.28 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 |
| DNS | 0.40 | 0.02 | 0.41 | 0.77 | 0.12 | 0.97 | 0.36 | 0.60 | 0.86 | 0.16 | 0.59 | 0.55 | 0.13 | 0.34 | 0.04 |
| Process | 0.44 | 0.82 | 0.64 | 0.96 | 0.74 | 0.03 | 0.31 | 0.08 | 0.08 | 0.05 | 0.08 | 0.43 | 0.17 | 0.06 | 0.01 |

The $p$-values are rounded to two decimal places, with rounded values at most 0.01 highlighted in green and values between 0.02 and 0.05 highlighted in yellow levels as the granularity parameter $\delta$ varies are more difficult to interpret. Even for a fixed window length and modality, the significance levels neither consistently increase nor decrease in $\delta$.

One plausible hypothesis for this experiments sensitivity to $\delta$ is that RH distance values exhibit periodic behavior both within and across modalities, reflecting the natural circadian rhythms one might expect from temporal cyber data. If this were the case, the choice of $\delta$ may skew the RH values sampled when constructing the representative before and after distributions. To check whether such periodicity is indeed present in the RH distance measurements on LANL, we constructed heatmaps of RH distances between all pairs of graphs over given time windows. As this requires a quadratic number of comparisons, it is worth noting this analysis is crucially facilitated by the lightweight computational complexity of RH distance. We examined heatmaps not only for windows surrounding anomalies, but also for time windows away from red team events. Figure 2 (left column) presents sample heatmaps for the Authentication, Flow and Process modalities spanning a 2-h time period. In an effort to select a representative window for short-term nominal RH behavior within each modality, this time period was selected so as to not include any red team events nor be preceded or followed by any red team events for 20 h. It is also worth pointing out that the RH distance between pairs of flow graphs regularly exceeds one, indicating that the rough guide for detecting anomalous behavior given in Simpson et al. (2015) is inappropriate for the cyber-security context. We also transformed each heatmap of pairwise RH distance values into a similarity matrix by applying the Gaussian kernel with $\sigma = 1$, and performed normalized Laplacian spectral clustering[3], as described by Ng et al. (2002). Under the corresponding heatmap, Fig. 2 (right column) plots the pairs of graphs belonging to common clusters, using a different color for each cluster (and white for different clusters).

A cursory examination of the heatmaps and their clustering suggests that the RH distance values for a given modality exhibit persistent and striking periodic patterns. Furthermore, in comparing the plots for Flow, Authentication, and Process, the differences in the periodic behavior of RH values are also apparent. While these periodicities are not entirely unexpected they are likely network- and data-dependent. Detecting and visualizing periodic behavior in network data is an active area of research, e.g., (Gove and Deason 2018; Hubballi and Goyal 2013; Price-Williams et al. 2017). As a consequence of this experiment and examination of the heatmaps in Fig. 2, it is clear that a single choice of granularity parameter $\delta$ is likely insufficient in establishing a representative distribution of RH values within a time window for any given modality. Accordingly, next we refine our experiment to better account for the inherent *multi-scale* and *multi-modal* nature of the LANL data.

One of the many difficulties with investigating the effectiveness of RH distance in detecting anomalous behavior associated with red team events in the LANL data sets is that the red team process is inherently multi-modal and multi-scale. That is, the red team events identified in the data are simply the first step of the red team intrusion process which could potentially affect all of the data modalities (process, DNS, flow, and authentication) and occur over multiple time scales. To attempt to deal with this issue, we craft an indicator for each time that considers the RH distance between graphs with multiple

[3] The prescribed number of clusters was chosen to coincide with the first observed gap in Laplacian eigenvalues, as in von Luxburg (2007).

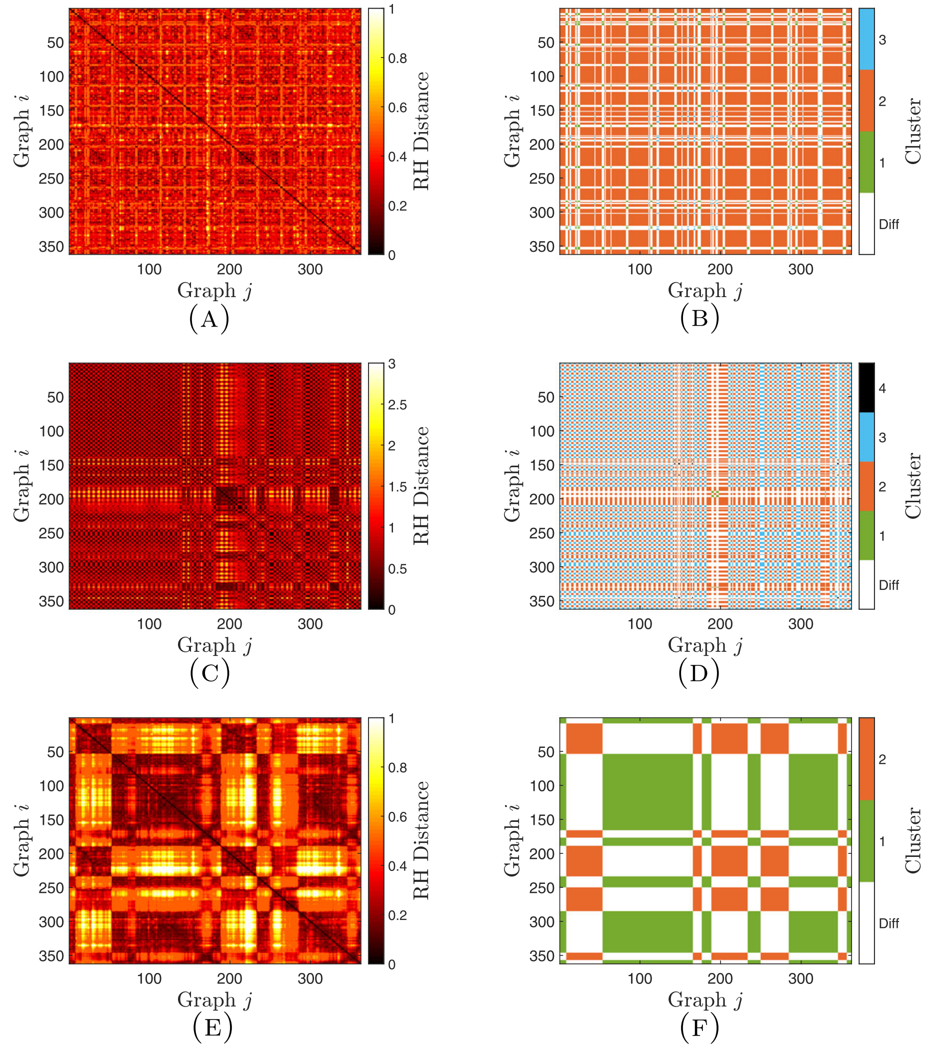


**Fig. 2** *Left Column:* RH distance heatmaps for nominal two-hour period. *Right Column:* Spectral clustering of similar graphs based on RH distances. **a** Authentication Heatmap. **b** Authentication Clustering. **c** Flow Heatmap. **d** Flow Clustering. **e** Process Heatmap. **f** Process Clustering

time differences and in multiple modalities. More concretely, let $\mathcal{S}$ be the set of potential data sources and let $\mathcal{G} = \{G_{t,s}\}$ be the collection of observed graphs indexed by the time $t$ and data source $s \in \mathcal{S}$. For any fixed timestamp $t$ and collection of differences $\mathcal{D}$, we will define the *profile vector* at time $t$, $v^{(t)}$, as the vector given by $(\mathcal{RH}(G_{t,s}, G_{t-\delta,s}))_{s\in\mathcal{S},\delta\in\mathcal{D}}$. Ideally, this profile could be used to aggregate the behavior across multiple modalities and multiple time scales and give a clearer picture of the overall state of system. However, there is a further complication with this approach in that the LANL data represents a system which has a naturally evolving behavior based on various temporal patterns of human activity (i.e. weekday vs. weeknight, circadian rhythms, etc.). To adjust for these temporal patterns, for every time $t$, source $s \in \mathcal{S}$, and difference $\delta \in \mathcal{D}$, we define a baseline behavior random variable $\mathcal{B}_{t,s,\delta}$ which is the random variable which represents the “typical” behavior of $\mathcal{RH}(G_{t,s}, G_{t-\delta,s})$. This baseline behavior can then be combined with the profile vector $v^{(t)}$ to generate a *temporal profile vector* $\hat{v}^{(t)}$ where for any $(s, \delta) \in \mathcal{S} \times \mathcal{D}$

we have $\hat{v}^{(t)}_{s,\delta} = \mathbb{P}\left(v^{(t)}_{s,\delta} - \epsilon < \mathcal{B}_{t,s,\delta} < v^{(t)}_{s,\delta} + \epsilon\right)$. We define the *temporal score* of the time $t$ as the geometric mean of the entries of $\hat{v}^{(t)}$ [4]. In what follows, we will calculate the temporal scores of time periods before and after a red team authentication event and show that there is a statistically significant difference between the behaviors. In fact, we will show that this temporal scoring methodology is more sensitive than using raw RH scores evaluated in Fig. 1 indicating that there are significant potential gains to be found by considering multi-modal and multi-scale indicators for anomalous behaviors.

Before applying our results to the LANL data sets, it remains to address how to estimate the distribution of the random variable $\mathcal{B}_{t,s,\delta}$ and how the $\epsilon$ term defines the temporal profile vector is chosen. For a fixed $t$ we estimate the empirical distribution for $\mathcal{B}_{t,s,\delta}$, we consider the RH distance between all pairs of graphs $(G_{t^*,s}, G_{t^*-\delta,s})$ where $t^*$ ranges over all times that differ from $t$ by a multiple of a week, plus or minus 10 min. In order to avoid biasing this empirical estimate we exclude times $t^*$ where there is a red team event in the interval $[\,t^*-\delta, t^*]$ as well as those that are within 10 min of $t$. As we see in Fig. 2 the typical variation of a RH distance changes significantly based on the modality of the observation, both in source and elapsed time between graphs. Thus, rather than fixing a particular value of $\epsilon$, we choose $\epsilon$ as one twentieth of the range of the empirical distribution for $\mathcal{B}_{t,s,\delta}$. Finally, for each of the 712 red team times provided in the LANL data we calculate the temporal scores for each graph in the 30 min before and after the red team time and apply the two-sample Kolmogorov-Smirnov test, see Table 2. We further segregate this data by whether or not additional red team events occur during the 30 min prior to the red team time.

It is clear from Table 2 that the temporal scores is far from a perfect indicator, as a non-negligible fraction of the changes associated with a red team event are not detected. Nonetheless, it is also apparent that for a relatively lightweight measure the RH distance exhibits reasonable effectiveness in distinguishing between nominal and anomalous behaviors. However, our conclusions must be somewhat tempered by the challenging nature of real world data and the LANL data in particular. Specifically, the lack of clear demarkation between anomalous and non-anomalous behavior as well as the limited time-scope of the investigation are significant caveats to any conclusions we make about the effectiveness of RH distance. In following section, we attempt to address these caveats by analyzing a synthetic temporal graph model inspired by the LANL cyber data.

## Simulated evolving networks

The study of temporal networks is concerned with the analysis and modeling of time-ordered sequences of graphs. In order to better understand temporal network dynamics, researchers have proposed a plethora of abstract models for their simulation (for a survey, see (Holme and Saramäki 2012)). In the present work, we consider a temporal graph model that belongs to the broader class of Markovian Evolving Graphs (MEGs) (Avin et al. 2008). Given a probability distribution over the set of all graphs on a fixed vertex set, MEGs have the defining property that the distribution at time $t$ is completely determined by that at $t-1$, thereby forming a sequence of random variables which satisfy the Markov property. Because of their generality and flexibility,

[4] As a practical matter, for those entries $(t, s, \delta)$ where there is insufficient or no data to estimate $\hat{v}^{(t)}_{s,\delta}$, the entry is dropped from the vector and ignored in the calculation of the temporal score.

**Table 2** Aggregate behavior of temporal scoring on a per event basis

| | $p \leq 0.10$ | $p \leq 0.05$ | $p \leq 0.01$ | Total |
|---|---|---|---|---|
| intervals with no prior red team events | 30 | 30 | 21 | 48 |
| intervals with prior red team events | 478 | 460 | 370 | 664 |

MEGs have been popularly used to study *information spreading processes*, such as file sharing on peer-to-peer networks, social network memes, and disease spreading (Clementi et al. 2014; Clementi et al. 2010).

In Hagberg et al. (2016) proposed a new MEG model, the design of which was informed by their study of LANL centralized authentication system cyber data (Kent 2014; 2016). In particular, they observed that these sequences of graphs exhibit certain stable global properties, such as skewed degree distributions, while local dynamics such as individual vertex neighborhoods change rapidly. To capture these dynamics, they designed a temporal model that can be used to preserve certain random graph structure while affording tunable control over the rate of dynamics. We refer to their model as the HLM model. Ultimately, Hagberg et. al utilized the HLM model to study *temporal reachability*; that is, the expected time (number of evolutions or transitions) before a constant fraction of the vertices are reachable from an arbitrary vertex. We note that although the HLM model was developed to capture abstract dynamics exhibited by cyber data, the HLM model need not be limited to simulating cyber phenomena. Although, the experiment that follows is driven by cyber-security structures and data, the is no a priori reason that a similar experiment could not be applied across the variety of domains for which the evolving nature of the HLM model is appropriate, such as communication networks, social networks, and (on a much slower time scale) transportation networks. In the remainder of this section, we study the sensitivity of RH distance in detecting several planted attack profiles, utilizing the HLM model to simulate the natural time evolution of a generic cyber network graph topology. Before describing our experimental methodology, we first begin by defining and briefly discussing the HLM model.

### Hagberg-Lemons-Mishra (HLM) model

As the HLM model can be viewed as a time-evolving generalization of the Chung-Lu model $\mathcal{G}(w)$, we will first briefly review the Chung-Lu model as introduced in Chung and Lu (2002); Chung and Lu (2004). The parameterization vector of the Chung-Lu model, $w$, is $n$-dimensional where $n$ is the number of vertices in the graph. Additionally, the vector $w$ satisfies that $w_v \leq \sqrt{\rho}$ for all $v$ where $\rho = \sum_{i=1}^{n} w_i$. From the parameter $w$ the Chung-Lu model is generated by including each edge $\{u, v\}$, independently, with probability $^{w_u w_v}/_{\rho}$. For overview of many of the known properties of the Chung-Lu model see the recent monograph (Chung and Lu 2006).

The HLM model generates an infinite sequence of graphs $G_0, G_1, G_2, \ldots$ with the property that there is a fixed vector $w$ such that for all $i$, $G_i \overset{\mathcal{D}}{=} \mathcal{G}(w)$ where $\overset{\mathcal{D}}{=}$ is equality in distribution. In order to generate this sequence an additional parameter, $\alpha$, is introduced to tune the extent to which graph $G_{i+1}$ is controlled by $G_i$. Specifically, $\alpha \in [0, 1]^n$ and $G_{i+1}$ is formed from $G_i$ by generating a masking set $M$ where each pair $\{u, v\}$ is in $M$ independently with probability $\sqrt{\alpha_u \alpha_v}$. For an edge $\{u, v\} \notin M$, $\{u, v\} \in G_{i+1}$ if and only if

$\{u, v\} \in G_i$, while each potential edge $\{u, v\}$ in $M$ is present independently with probability $w_u w_v/\rho$. In summary, we have that

$$\mathbb{P}(\{u,v\} \in G_{i+1}) = \begin{cases} 1 - \sqrt{\alpha_u \alpha_v} + \sqrt{\alpha_u \alpha_v} w_u w_v/\rho & \{u,v\} \in G_i \\ \sqrt{\alpha_u \alpha_v} w_u w_v/\rho & \{u,v\} \notin G_i \end{cases}.$$

The fact that $G_{i+1} \overset{\mathcal{D}}{=} G_i$ follows by induction and the observation that

$$w_u w_v/\rho(1 - \sqrt{\alpha_u \alpha_v} + \sqrt{\alpha_u \alpha_v} w_u w_v/\rho) + (1 - w_u w_v/\rho)\sqrt{\alpha_u \alpha_v} w_u w_v/\rho = w_u w_v/\rho.$$

We note that there is a natural trivial generalization of the HLM model where the edge probability $w_u w_v/\rho$ is replaced with arbitrary values in $p_{uv} \in [0, 1]$. In this case, at each time step the network is distributed over graphs like $\mathcal{G}(P)$, the generic independent edge graph model with parameter $P$. Similarly, the evolution parameter $\alpha$ can be generalized to a symmetric matrix $A \in [0, 1]^{n \times n}$. We note that several well studied models fall into this framework, including the stochastic block model, stochastic Kronecker graphs (Leskovec et al. 2005; Mahdian and Xu 2007), random dot product graphs (Young 2008; Young and Scheinerman 2008; 2007), and the inhomogeneous random graph model (Bollobás et al. 2007; Söderberg 2002). In order to maintain consistent notation, we will specify all of the experiments in this work in terms of this generalized HLM model even though most of generative matrices $P$ come from the Chung-Lu model. Further, with the aim of having the minimum number of free-parameters we will only consider HLM evolutions where $A_{uv} = A_{xy}$ for all $u \neq v$ and $x \neq y$. We will further slightly abuse notation and refer to this common value as $\alpha$.

Finally, we note that this generalized framework can be further expanded by allowing the parameter matrix $P$ to depend on the time step $t$. In particular, we have

$$\mathbb{P}(\{u,v\} \in G_{t+1} \mid G_t) = \begin{cases} (1-\alpha) + \alpha p_{uv}^{(t+1)} & \{u,v\} \in G_t \\ \alpha p^{(t+1)} & \{u,v\} \notin G_t \end{cases}.$$

It is worth mentioning that in this case $G_{t+1}$ is not distributed like $\mathcal{G}(P^{(t+1)})$ because of the possibility of edges being present from earlier timesteps. In fact, it is an easy exercise to show that the edges of $G_t$ are distributed according to $(1-\alpha)^t P^{(0)} + \sum_{i=1}^{t} \alpha(1-\alpha)^{t-i} P^{(i)}$.

## Experimental setup

In our experimental setup, in keeping with the lightweight nature of the RH calculation, we focus on the detection of small anomalies in extremely sparse graphs, such as we observed in small time windows for the LANL data set and other proprietary network flow data. For the sparse graphs we consider two different fixed degree distributions. Both of these degree distributions are formed by choosing 5000 samples from some fixed probability distribution. For the first degree distribution, we estimate a degree density function using a smoothed median estimator from a selection of one minute graphs in the LANL network flow data set, see Fig. 3a. The resulting degree density function results in a "power-law" like degree distribution with exponent approximately 3.5. Although the resulting degree density function is not truly a power-law distribution, we will abuse notation and refer to it as a "power-law" degree distribution. For a discussion of the difficulties and appropriateness of the power-law degree distribution for real data the interested reader is referred to the recent work (Broido and Clauset 2018). The resulting distribution has 4742 edges in expectation as well as maximum expected degree 961. The second

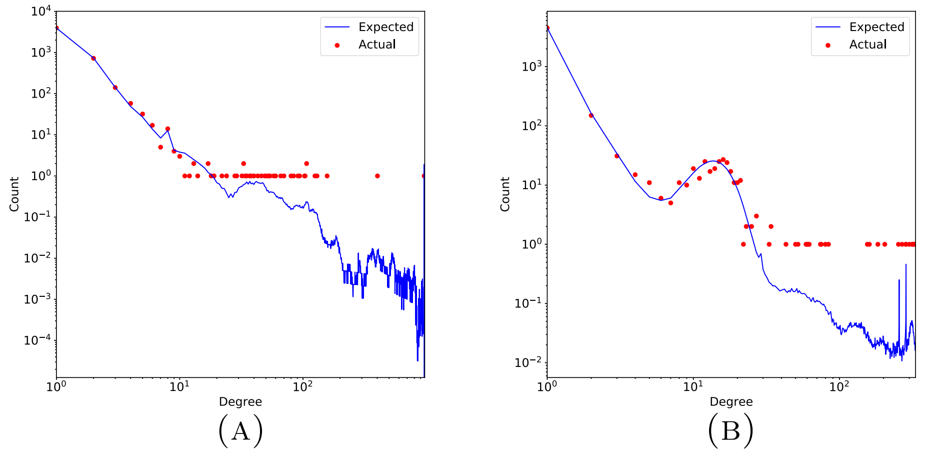


**Fig. 3** log-log Degree Distribution. **a** Power-Law. **b** Bump Power-Law

distribution represents what we call a "bump power-law," that is, a power-law distribution coupled with an approximately binomially distributed "bump" at higher degrees. This can be thought of as a more hub-and-spoke style network where the degree of the spoke vertices are approximately power-law distributed while the degree of the hub vertices are approximately binomially distributed. For the bump power-law distribution, the degree probabilities were explicitly estimated from a collection of several thousand graphs generated from a proprietary enterprise boundary network flow data set (see Fig. 3b). This resulting distribution has 6067 edges in expectation as well as maximum expected degree 327.[5]

We will also consider two different styles of anomalies involving between 10 and 50 edges. The first anomaly involves three randomly chosen vertices adding some number of edges to the rest of the network uniformly at random. We view this as behavior consistent with a probe or scan of the network structure. For the second anomaly a random collection of vertices are chosen and a random spanning tree is added among those vertices. We view this as behavior consistent with lateral movement scenario where an attacker is exploring the network by moving from machine to machine. They may backtrack and try different routes (thus a tree rather than just a path) as needed.

For each of these 420 scenarios (two different degree distributions, two different anomaly types, five different anomaly sizes, and 21 different values of $\alpha$) we produce 1000 different pairs of graphs $(G, G')$ where $G$ is a random instance of the Chung-Lu model with the chosen degree distribution and $G'$ is formed from $G$ by performing one step of the HLM evolution with the chosen parameter $\alpha$, and then adding a random instance of the chosen anomaly of the chosen size. In this way, the anomaly occurs concurrently with the natural evolution of the network, as might be typical of real-world data. For each of our 420 scenarios, and 1000 pairs of graphs within the scenario, we compute the RH distance between $G$ and $G'$ to get a distribution of RH distances for the anomalous transition.

[5] It is worth mentioning that both of these degree distributions violate the standard assumption for the Chung-Lu model $\max_v w_v \leq \sqrt{\rho}$. To deal with this, we replace that edge probabilities of $w_u w_v/\rho$ with $\min\{1, w_u w_v/\rho\}$. However, as there are under 200 pairs $\{u, v\}$ where $w_u w_v > \rho$ for each of the degree distributions, this makes a minimal difference in the model.

### Anomalous versus nominal relative hausdorff distance

In this section we consider whether the anomalous transitions in the HLM model result in a different distribution of RH distances than a nominal transition. To this end, for each degree distribution and choice of $\alpha$, we simulate 10,000 different HLM transitions to develop a baseline distribution of RH distances, see Fig. 4. For each of the 420 anomalous scenarios we calculate the 2-sample Kolmogorov-Smirnov $p$-value (Young 1977) between the previously calculated anomalous distribution and this baseline distribution. For each of the 420 different anomaly scenarios the KS test significance value is less that 0.01, indicating that we can reject the null hypothesis that the distribution of RH distances for an anomalous HLM transition is the same as the distribution for non-anomalous transitions. In particular, this means that in a statistical sense the RH distance is able to pick up on anomalous evolution of the degree distribution, even when the anomaly only consists of 10 edges.[6] In the next subsection we will consider the effectiveness of the RH distance in detecting anomalous behavior directly, rather than statistically.

### Anomaly detection

In this section we consider how RH distance could be used to detect anomalous behavior in a streaming environment and compare to the performance with a similarly lightweight measure (KS distance) as well as a "ideal" measure (graph edit distance). To compare between these three methods in a non-parametric way (i.e. without introducing a "anomaly threshold") we introduce the idea of an *anomaly score* of an observation with respect to a theoretically or empirically observed baseline distribution. Specifically, let the random variable $Z$ have theoretical or empirical cumulative distribution function $f_Z\colon \mathbb{R} \to [0,1]$. We will then say that a particular observation $z$ (not necessarily distributed as $Z$) has an anomaly score relative to $f_Z$ of $2\left\| f_Z(z) - \frac{1}{2} \right\|$. Note that this score takes on values from $[0,1]$ with values closer to one being more "anomalous." This score can be thought of as measuring the deviation of the observation $z$ from the bulk of the distribution of $Z$.

Before turning to a direct comparison between anomaly scores for RH distance, KS $p$-values[7], and edit distance, we consider the performance of each of these anomaly scores in isolation via ROC-like curves, presented for a subset of our 20 scenarios (2 distributions, 2 anomaly types, 5 levels of each anomaly) in Figs. 5, 6, 7, and 8. Note that as the relative frequency of anomalous and non-anomalous behavior is unknown, these are not truly ROC curves but rather implicit plots $(x(t), y(t))$ where $t$ is some threshold value. Specifically, $y(t)$ is the fraction of the anomalous transitions that have anomaly score at least $t$, i.e. "true positives", where $x(t)$ is the fraction of non-anomalous transitions that have anomaly score at least $t$, i.e. "false positives". At this point it is worth pointing out that if $z$ is identically distributed with the random variable $Z$, then the anomaly score for $z$ is uniformly distributed over $[0,1]$. As a consequence, we can explicitly define $x(t) = 1 - t$. To compute the ROC curve for one scenario we used the previously computed cumulative distribution function for the 10,000 non-anomalous transitions as $f_Z$. Then, for each

[6] It is important to note that this is not always the case. For example, in an experiment that is not reported for space limitations, we synthetically generated a degree distribution with a power-law exponent of 4 and average degree around 1.4, the anomalies resulted in a range of KS statistics including several scenarios which were not statistically distinguishable.

[7] From this point on in this work, although we will be using the KS $p$-value we will be treating it simply as a distance measure rather than a statistical quantity. In particular, we will make no assumptions about the meaning of large or small values of the $p$-value other than as a means of measuring the "closeness" between two degree distributions.

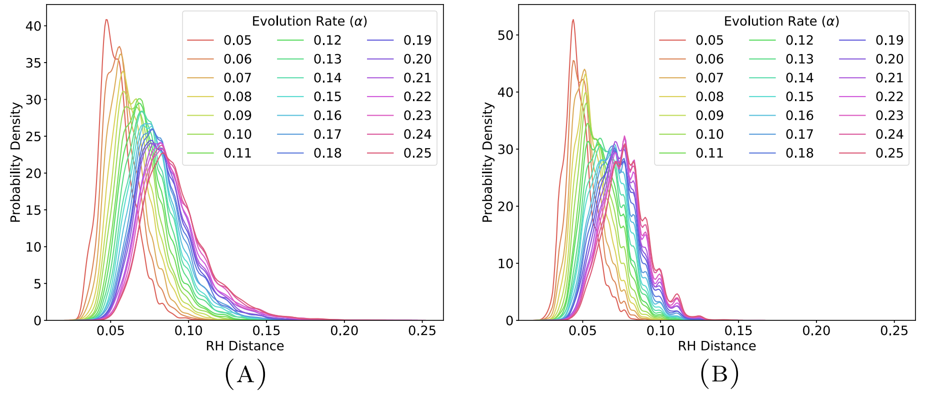


**Fig. 4** Distribution of RH distance under HLM evolution. **a** Power-Law. **b** Bump Power-Law

of the 1000 anomalous transitions we use the RH distance as $z$ and compute the anomaly score for that value in the context of $f_Z$.

Overall we can see that, for detecting anomalies, edit distance would be preferred to RH distance, which would in turn be preferred to the KS statistic. However, for the bump power-law, under the lateral movement anomaly with 10 edges, we see that the RH distance outperforms the edit distance, see Fig. 7.

### *Kolmogorov-Smirnov (KS)*

For this section we will compare the anomaly score of the RH distance with the anomaly score of the KS $p$-value (significance value) between successive degree distributions both for the 420 anomalous scenarios and the 40 baseline distributions. We note that since the degree distributions are discrete valued, the application of the KS test for hypothesis testing is not necessarily appropriate, however as we are interested in statistical behavior of the significance test and KS is widely used in the network analysis literature (see, for instance, (Aliakbary et al. 2014; Broido and Clauset 2018; Simpson et al. 2015)) we will ignore these technical issues.

Figure 9 gives the relative performance of the KS and RH anomaly scores across all 420 anomaly scenarios. The $y$ value counts how many of the 1000 cases RH outperforms KS. We can see that the RH distance outperforms KS the most for the scenario where there is a 50-edge scan anomaly on the power-law distribution with an evolution rate of 0.23. In this case, the RH anomaly score is larger in 924 of the 1000 different trials. We see that overall,

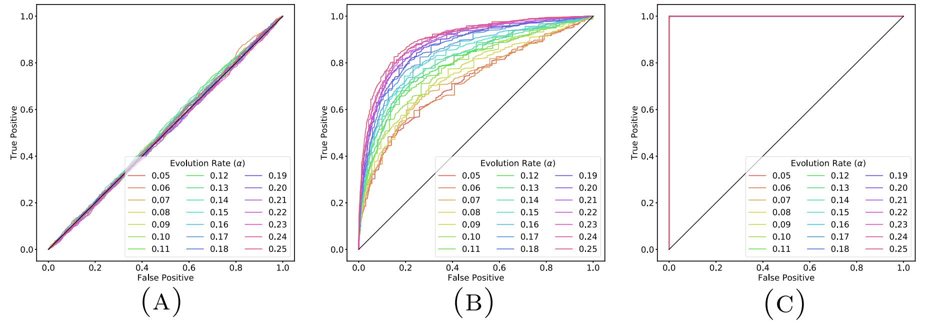


**Fig. 5** Power-Law Distribution, Scan, 10 edges. **a** Kolmogorov-Smirnov. **b** Relative Hausdorff. **c** Edit

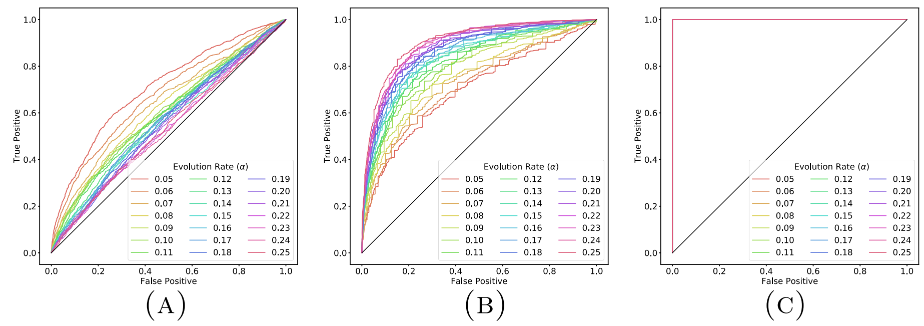


**Fig. 6** Power-Law, Lateral Movement, 50 edges. **a** Kolmogorov-Smirnov. **b** Relative Hausdorff. **c** Edit

excepting cases with a low-evolution rate and larger, lateral movement anomalies, RH distance is clearly superior, especially for the power-law degree distribution. It is worth noting that relative performance of the KS statistic improves when considering the lateral movement anomaly rather than the scan anomaly. Since the degree change caused by lateral movement is spread across many vertices (as opposed to scan where the primary change is spread across only three vertices), this result can be explained by the well known sensitivity of the KS test to variation away from the a tails of the distribution (Simpson et al. 2015).

We note that a direct binary comparison between the two measures may not tell the whole story of their relative performance. For instance, in an extreme case, one can imagine one of the two measures taking on a fixed large value (indicating an anomaly) while the other takes on both small values and, more frequently, a value that is slightly larger than the other anomaly score. We separate the anomalous pairs of graphs into two sets according to whether their RH or KS anomaly score is higher. Then, for each of these two classes we report in Fig. 10 (across all 420 anomaly scenarios) the mean difference between the scores with error bars representing one standard deviation of range around this mean. We note that the RH anomaly score typically exceeds the KS anomaly score by about 0.4, while the KS anomaly score typically exceeds the RH anomaly score by between 0.2 and 0.3. Further, the standard deviation across all cases the average gap between the RH and KS anomaly score is fairly consistently in the range [ 0.2, 0.3] essentially independent of all parameters. Together, the data in Figs. 9 and 10, indicates that the RH distance

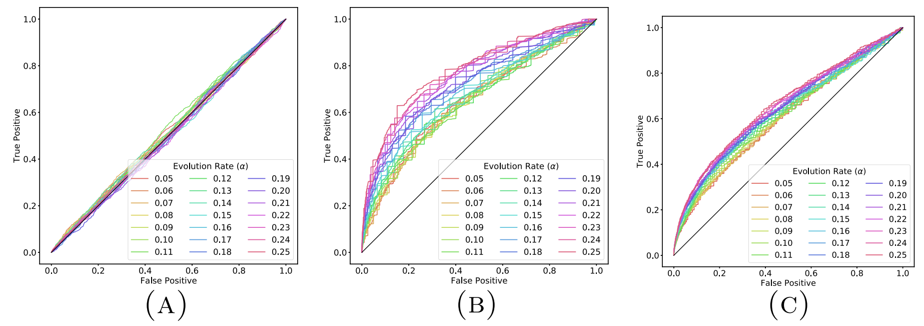


**Fig. 7** Bump Power-Law, Lateral Movement, 10 edges. **a** Kolmogorov-Smirnov. **b** Relative Hausdorff. **c** Edit

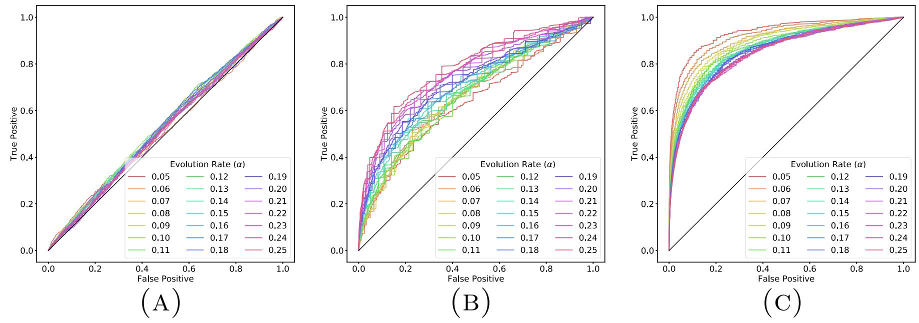


**Fig. 8** Bump Power-Law, Scan, 50 edges. **a** Kolmogorov-Smirnov. **b** Relative Hausdorff. **c** Edit

is significantly more sensitive than KS distance to the broad range of anomalies we have investigated.

### *Graph edit distance*

In this section we compare the sensitivity of RH distance to a "perfect information" aggregate measure, in particular graph edit distance. Recall that the edit distance between two graphs $F$ and $G$ is the minimum "weight" of a sequence of edge/vertex additions/deletions needed to transform $F$ to $G$. In general this quantity is $\mathcal{NP}$-complete to compute (see (Zeng et al. 2009)) and likely impractical to even approximate (Lin 1994). This complexity is driven by the difficulty in finding the optimal alignment between the vertices of $F$ and $G$ which maximizes the edge overlap between $F$ and $G$. For the HLM model, this problem is mitigated by the natural alignment between the graphs generated at consecutive time steps. Thus, for purposes of this section we will approximate the graph edit distance as the number of edges that "flip" during each evolution of the HLM model.

The following lemma allows us to significantly simplify the calculation of the anomaly score for edit distance by approximating the baseline distribution with the large $n$ limit.

**Lemma 1** *Let $G$ be an random graph distributed according to $\mathcal{G}(P)$ and let $G'$ be the graph formed by one iteration of the Hagberg-Lemons-Mishra evolution with evolution*

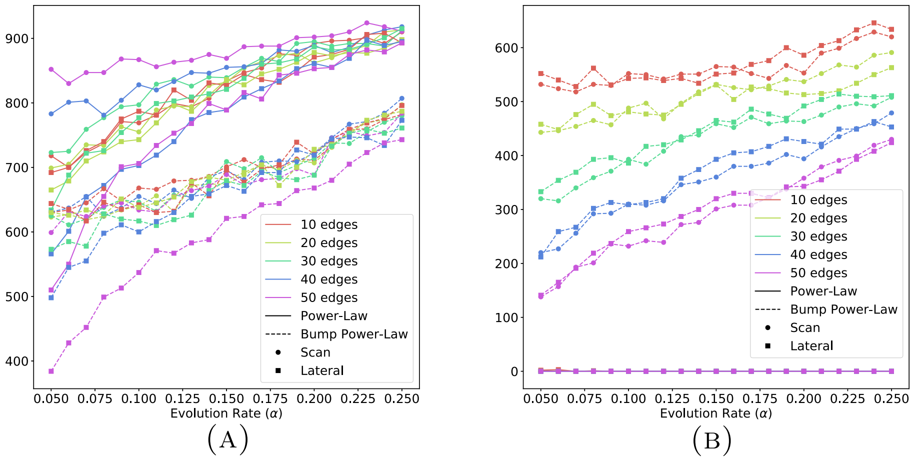
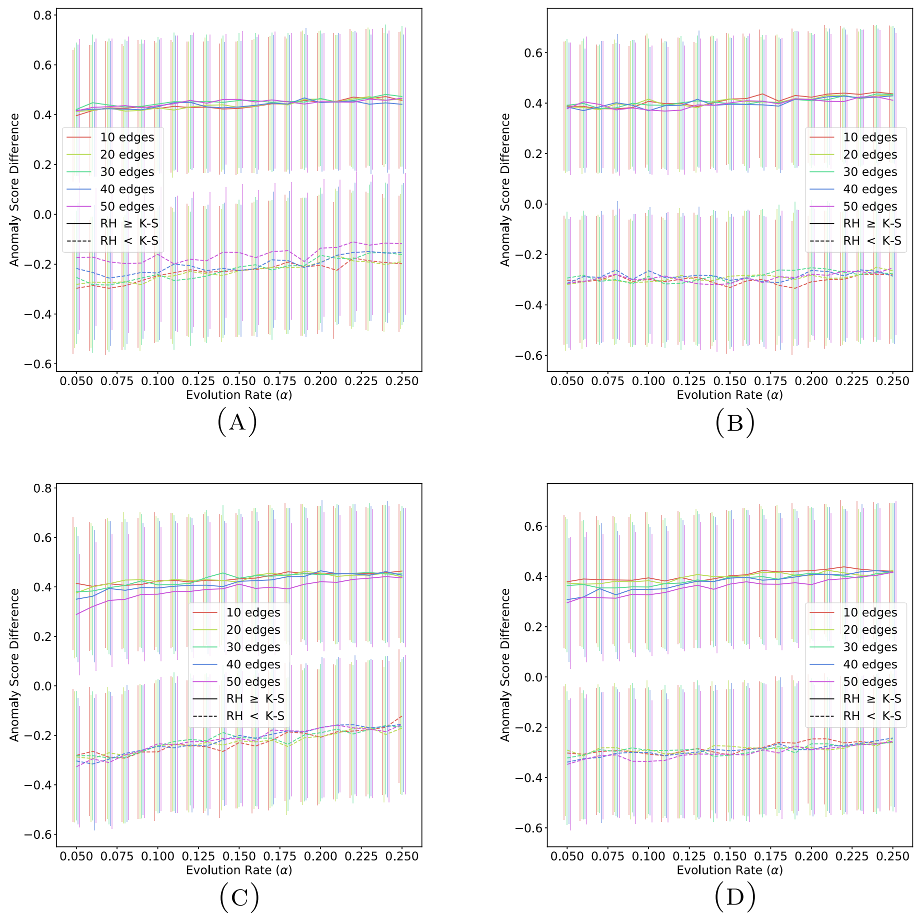


**Fig. 9** Relative Performance of RH and KS Anomaly Scores. **a** RH versus KS. **b** RH versus Edit. **a** RH versus KS. **b** RH versus Edit



(A) (B) (C) (D)

**Fig. 10** Difference between KS anomaly scores and RH anomaly scores. **a** Power-Law Scan. **b** Bump Power-Law Scan. **c** Power-Law Lateral movement. **d** Bump Power-Law Lateral movement

*parameter $\alpha$ and probability matrix $P'$. Let $X$ be the random variable that counts the number of edges that differ between $G$ and $G'$. If $\mathrm{Var}(X) \to \infty$, then $X$ is asymptotically normally distributed.*

*Proof* Let $X_{ij}$ be the indicator function for the random variable that edge $\{i,j\}$ is present in precisely one of $G'$ and $G$ and observe that $X = \sum_{i<j} X_{ij}$. We recall that by the Lyapunov Central Limit Theorem ((Billingsley 2008), p. 362), we have that

$$\frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

if there is some $\delta > 0$ such that

$$\lim_{n\to\infty} \frac{1}{\mathrm{Var}(X)^{2+\delta/2}} \sum_{i<j} \mathbb{E}\left[\left|X_{ij} - \mathbb{E}[X_{ij}]\right|^{2+\delta}\right] = 0.$$

Fixing $\delta = 1$, we note that

$$\begin{aligned}\sum_{i<j} \mathbb{E}\left[|X_{ij} - \mathbb{E}[X_{ij}]\,|^3\right] &= \sum_{i<j} \mathbb{E}[X_{ij}]\,(1-\mathbb{E}[X_{ij}]\,)^3 + (1-\mathbb{E}[X_{ij}]\,)\mathbb{E}[X_{ij}]^3 \\ &= \sum_{i<j} \mathbb{E}[X_{ij}]\,(1-\mathbb{E}[X_{ij}]\,)\left(\mathbb{E}[X_{ij}]^2 + (1-\mathbb{E}[X_{ij}]\,)^2\right) \\ &\leq \sum_{i<j} \mathbb{E}[X_{ij}]\,(1-\mathbb{E}[X_{ij}]\,) \\ &= \mathrm{Var}(X).\end{aligned}$$

Thus, if $\mathrm{Var}(X) \to \infty$, then

$$\lim_{n\to\infty} \frac{1}{\mathrm{Var}(X)^{3/2}} \sum_{i<j} \mathbb{E}\left[|X_{ij} - \mathbb{E}[X_{ij}]\,|^3\right] = 0$$

and $X$ is normally distributed. □

It is worth mentioning that this, in principle, allows for an explicit formula for the distribution of the anomaly score for edit distance in a wide range of baseline and anomalous behaviors, namely

$$\mathbb{P}(S \leq s) = \Phi\left(\frac{\mu-\mu_A}{\sigma_A} + \frac{\sigma_A}{\sigma}\Phi^{-1}\left(\frac{s+1}{2}\right)\right) - \Phi\left(\frac{\mu-\mu_A}{\sigma_A} - \frac{\sigma_A}{\sigma}\Phi^{-1}\left(\frac{s+1}{2}\right)\right),$$

where $(\mu, \sigma)$ and $(\mu_A, \sigma_A)$ are the mean and distribution of the baseline and anomalous evolutions, respectively, and $\Phi$ is the cumulative distribution function of the standard normal distribution. However, given the correlated nature of the anomalies the calculation of $\sigma_A$ is tedious, so we will empirically estimate this distribution.

Figure 9b again presents the relative performance of the anomaly scores, this time for edit distance and RH distance, for all 420 anomaly trials. Again the $y$ value counts how many of the 1000 cases RH outperforms edit distance. We note that in the best case (bump power-law degree distribution, lateral movement anomaly, 10 edges, $\alpha = 0.24$), the RH distance anomaly score is larger than the edit distance anomaly scores 646 times. However, for the power-law case the RH anomaly scores essentially never outperform the edit distance anomaly scores. This failure is mitigated by the fact that, as mentioned earlier, in many cases the edit distance is computationally infeasible, while the RH distance requires minimal computational overhead. It is also worth mentioning that we can see a clear degradation of performance for edit distance as the size of the anomaly decreases and the evolution rate increases. This phenomenon can be explained by observing that the anomaly score for edit distance is driven by a $z$-score of the anomaly, which is linearly correlated with the anomaly size and inversely correlated with the standard deviation of the baseline distribution. Additionally, the variance baseline distribution of edit distance is linear related to the evolution rate, resulting in significantly decreased sensitivity at high evolution rates.

We further compare the relative behavior of the edit distance anomaly scores and the RH distance anomaly scores, in the same way as we did for KS above, by considering the average difference between the anomaly scores in the cases where the RH anomaly score is larger (positive values) and in the case the edit distance anomaly score is larger (negative values). As the RH distance anomaly score is essentially never larger than the edit distance anomaly score for the power-law distribution, we restrict our attention here to the bump power-law distribution. In Fig. 11, we again report the relative magnitude of the differences with the error bars representing an interval one standard deviation away from the mean. Again we can see a clear stratification of the behavior with the RH anomaly scores performing better as the size of the anomaly decreases. We also note the mild improvement in the performance of RH distance as the evolution rate increases, likely reflecting the decreased sensitivity of edit distance (due to larger variance).

Interestingly, the standard deviation is essentially constant over all choices of degree distribution, anomaly type, anomaly size, and evolution rate and is also roughly equal to the standard deviations shown in Fig. 10. Furthermore, the magnitude of the standard deviation is close to the minimal possible standard deviation given by the generalization of Bhatia-Davis inequality for the variance of a bounded random variable (Agarwal et al. 2005). As the extremal distribution is given by point masses at the end points of the distribution, this indicates that there are three essentially distinct outcomes: the RH distance anomaly score is significantly larger than the edit distance anomaly score, the RH and edit distance anomaly scores are essentially the same, and the edit distance anomaly score is significantly larger than the RH distance anomaly score. Furthermore, this holds regardless of the size and nature of anomaly or evolution rate and also holds when replacing edit distance with KS distance (for both degree distributions).

## Conclusions

In this work, we conducted an experimental and statistical study of Relative Hausdorff distance in the context of time-evolving sequences of graphs. Applying RH distance as an anomaly detection tool, we first tested its detection of red team events across multiple modalities in real cyber security data. We found evidence that RH distance values register statistical change-points at red team events, although these results were sensitive to window length, the granularity of pairwise RH measurements, and subject to the limitations of the data. In order to test RH distance in a more controlled and rigorous manner, we then turned our attention to a temporal graph model inspired by cyber-data.

Using this temporal graph model to generate synthetic sequences of evolving graphs, we experimentally tested the sensitivity of RH distance to two attack profiles. To broaden our tests scope, we considered a multitude of parameter settings in which we varied the input degree distribution, temporal evolution rate, and intensity of the attack signal. In

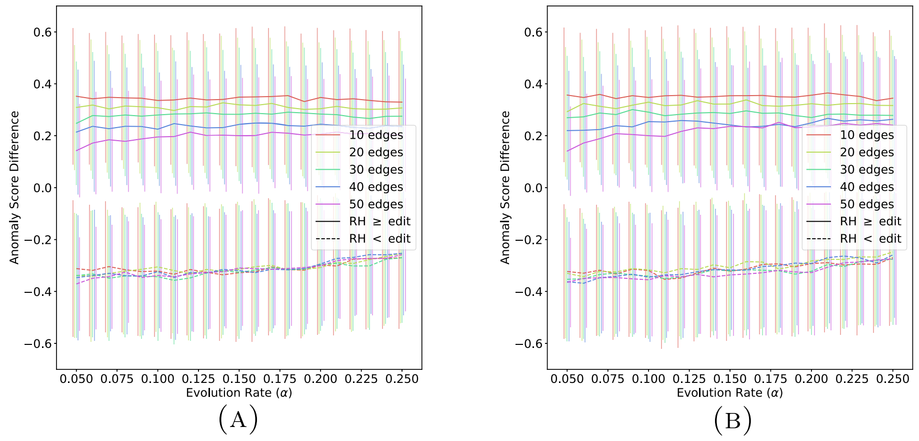


**Fig. 11** Difference between edit distance and RH distance. **a** Bump Power-Law Scan. **b** Bump Power-Law Lateral movement

its own right, RH distance performed respectably, yielding ROC curves above the line of no-discrimination for every scenario tested. Compared with other similarity measures, RH distance consistently outperformed another lightweight similarity measure based on Kolmogorov-Smirnov distance, while its performance against the computationally-intensive edit distance was more mixed: while edit distance clearly outperformed RH distance under scenarios featuring the power law degree distribution, RH distance was better able to detect the low-intensity lateral movement attack under the bump power law degree distribution.

Anomaly detection generally, and even specifically in cyber security, is not amenable to a "one method to rule them all" mentality. Indeed, there are many types of anomalies and one does not expect them to all be caught by the same detector. It is important to recognize that our analysis does not use all of the available information pertinent to real cyber data. Distilling a time interval of data down to a single graph and removing all metadata is likely to introduce many false positives. It could be that a graph is anomalous given the recent context, but the behavior is fully expected by cyber security operations analysts (e.g. a daily backup may appear to be an exfiltration if the IP addresses involved aren't considered). In the other direction, if the graph does not contain the metadata that would flag an anomaly, this may similarly introduce false negatives. By integrating metadata into our analysis, it is possible that as anomalies are discovered, this metadata could be used to help classify them as benign and nefarious anomalies. Lastly, it is also worth noting that our analyses considered the entire data in a given period, as opposed to an online approach. Whether and how RH distance might be utilized in online anomaly detection frameworks remains another open topic for future research.

### Abbreviations

RH: Relative Hausdorff; KS: Kolmogorov-Smirnov; GED: Graph edit distance; HLM: Hagberg-Lemons-Mishra; MEG: Markovian evolving graph

### Acknowledgements

The authors would like to thank our colleague Carlos Ortiz-Marrero for reviewing the manuscript.

### Authors' contributions

EHP and SGA conducted the literature review, KEN processed the LANL data, SGA and SJY implemented the experiments in "Los Alamos National Laboratory (LANL) cybersecurity data" section, SJY implemented the experiments in "Simulated evolving networks" section and proved Lemma 1. All authors jointly conceived and wrote the paper, designed and analyzed the experiments, and approved the final manuscript. At the preference of all authors, the authors are listed in alphabetical order. All authors read and approved the final manuscript.

### Funding

PNNL is operated by Battelle for the United States Department of Energy under Contract DE-AC05-76RL01830. The design of the study and all data analysis and interpretation was done under the direction of the authors of this paper.

### Availability of data and materials

The data used in "Los Alamos National Laboratory (LANL) cybersecurity data" section is available at https://csr.lanl.gov/data/cyber1/. The proprietary enterprise boundary network flow data used to estimate the bump power-law distribution featured in "Simulated evolving networks" section is unavailable. The remaining data and code will be made available upon request to the corresponding author, subject to institutional approval.

### Competing interests

The authors have no competing interests to declare.

### Author details

[1] Pacific Northwest National Laboratory, Richland, WA 99352, United States. [2] Pacific Northwest National Laboratory, Seattle, WA 98109, United States.

Received: 15 March 2019 Accepted: 23 August 2019
Published online: 17 October 2019

## References

Agarwal PK, Fox K, Nath A, Sidiropoulos A, Wang Y (2018) Computing the gromov-hausdorff distance for metric trees. ACM Trans Algoritm 14:1–20

Agarwal R, Barnett NS, Cerone P, Dragomir SS (2005) A survey on some inequalities for expectation and variance. Comput Math Appl 49:429–480

Aggarwal CC, Zhao Y, Philip SY (2011) Outlier detection in graph streams. IEEE. https://doi.org/10.1109/icde.2011.5767885

Akoglu L, Faloutsos C (2010) Event detection in time series of mobile communication graphs. In: 27th Army science conference, Orlando. pp 77–79

Akoglu L, Tong H, Koutra D (2014) Graph based anomaly detection and description: a survey. Data Min. Knowl. Discov. 29:626–688

Aksoy S, Nowak K, Young S (2018) A linear-time algorithm and analysis of graph relative hausdorff distance. in preprint. 1906.04936

Aliakbary S, Habibi J, Movaghar A (2014) Quantification and comparison of degree distributions in complex networks. In: 7'th International Symposium on Telecommunications (IST'2014). IEEE. pp 464–469. https://doi.org/10.1109/istel.2014.7000748

Avin C, Koucký M, Lotker Z (2008) How to explore a fast-changing world (cover time of a simple random walk on evolving graphs). In: Automata, Languages and Programming. Springer, Berlin, Heidelberg. pp 121–132

Banič I, Taranenko A (2015) Measuring closeness of graphs—the hausdorff distance. Bull Malays Math Sci Soc 40:75–95

Billingsley P (2008) Probability and measure. Wiley, Hoboken

Blondel VD, Gajardo A, Heymans M, Senellart P, Dooren PV (2004) A measure of similarity between graph vertices: Applications to synonym extraction and web searching. SIAM Rev 46:647–666

Bollobás B, Janson S, Riordan O (2007) The phase transition in inhomogeneous random graphs. Random Struct Algoritm 31:3–122

Broido AD, Clauset A (2018) Scale-free networks are rare. arXiv preprint. arXiv:1801.03400

Chen P, Choudhury S, Hero AO (2016) Multi-centrality graph spectral decompositions and their application to cyber intrusion detection. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. pp 4553–4557. https://doi.org/10.1109/icassp.2016.7472539

Choi J (2019) Gromov-hausdorff distance between metric graphs. https://math.mit.edu/research/highschool/primes/materials/2018/Choi.pdf

Chung F, Lu L (2002) The average distances in random graphs with given expected degrees. Proc Natl Acad Sci 99:15879–15882

Chung F, Lu L (2004) The average distance in a random graph with given expected degrees. Internet Math 1:91–113

Chung F, Lu L (2006) Complex graphs and networks, vol. 107 of CBMS Regional Conference Series in Mathematics. Published for the Conference Board of the Mathematical Sciences, Washington, DC

Clementi A, Silvestri R, Trevisan L (2014) Information spreading in dynamic graphs. Distrib Comput 28:55–73

Clementi AEF, Macci C, Monti A, Pasquale F, Silvestri R (2010) Flooding time of edge-markovian evolving graphs. SIAM J Discrete Math 24:1694–1712

Edwards DA (1975) The structure of superspace. In: Studies in Topology. Elsevier. pp 121–133. https://doi.org/10.1016/b978-0-12-663450-1.50017-7

Fernández M-L, Valiente G (2001) A graph distance metric combining maximum common subgraph and minimum common supergraph. Pattern Recogn Lett 22:753–758

François J, Wang S, Engel T, et al. (2011) Bottrack: tracking botnets using netflow and pagerank. In: International Conference on Research in Networking. Springer. pp 1–14

Gao X, Xiao B, Tao D, Li X (2009) A survey of graph edit distance. Pattern Anal Appl 13:113–129

Gaston ME, Kraetzl M, Wallis WD (2006) Using graph diameter for change detection in dynamic networks. Australas J Comb 35:299–312

Gibbons JD, Chakraborti S (2011) Nonparametric statistical inference. Springer, New York

Giuseppe J, Roberto V, Cesare F (2011) An introduction to spectral distances in networks. Front Artif Intell Appl 226:227–234

Gove R, Deason L (2018) Visualizing automatically detected periodic network activity. In: Proceedings of the IEEE Symposium on Visualization for Cyber Security. Center for Open Science. https://doi.org/10.31219/osf.io/xpwfe

Gromov M (1981) Structures métriques pour les variétés riemanniennes. Textes Math Matiques Math Texts 1:iv+152

Hagberg A, Lemons N, Misra S (2016) Temporal reachability in dynamic networks. In: Dynamic Networks and Cyber-Security, WORLD SCIENTIFIC (EUROPE). WORLD SCIENTIFIC (EUROPE). pp 181–208. https://doi.org/10.1142/9781786340757_0009

Hausdorff F (1914) Grundzuge der Mengenlehre. Am Math Soc. Leipzig: Veit, ISBN 978-0-8284-0061-9 Reprinted by Chelsea in 1949

Holme P, Saramäki J (2012) Temporal networks. Phys Rep 519:97–125

Hubballi N, Goyal D (2013) Flowsummary: Summarizing network flows for communication periodicity detection. In: International Conference on Pattern Recognition and Machine Intelligence. Springer. pp 695–700. https://doi.org/10.1007/978-3-642-45062-4_98

Ishibashi K, Kondoh T, Harada S, Mori T, Kawahara R, Asano S (2010) Detecting anomalous traffic using communication graphs. In: Telecommunications: The Infrastructure for the 21st Century (WTC), 2010. VDE, Berlin. pp 1–6

Kent A (2014) Anonymized user-computer authentication associations in time, tech. report. Los Alamos National Lab.(LANL), Los Alamos

Kent AD (2015) Comprehensive, Multi-Source Cyber-Security Events. Los Alamos National Laboratory, London

Kent AD (2015) Cybersecurity Data Sources for Dynamic Network Research. In: Dynamic Networks in Cybersecurity. Imperial College Press

Kent AD (2016) Cyber security data sources for dynamic network research. In: Dynamic Networks and Cyber-Security. World Scientific, Singapore. pp 37–65

Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J. ACM 46:604–632

Lee H, Chung MK, Kang H, Kim B-N, Lee DS (2011) Computing the shape of brain networks using graph filtration and gromov-hausdorff metric. In: Lecture Notes in Computer Science. Springer, Berlin Heidelberg. pp 302–309
Leskovec J, Chakrabarti D, Kleinberg J, Faloutsos C (2005) Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In: Knowledge Discovery in Databases: PKDD 2005. Springer, Berlin Heidelberg. pp 133–145
Lin CL (1994) Hardness of approximating graph transformation problem. In: Algorithms and Computation. Springer, Berlin Heidelberg. pp 74–82
Mahdian M, Xu Y (2007) Stochastic kronecker graphs. In: International workshop on algorithms and models for the web-graph. Springer, Berlin Heidelberg. pp 179–186
Marsaglia G, Tsang WW, Wang J (2003) Evaluating kolmogorov's distribution. J. Stat. Softw. 8:1–4
Matulef KM (2017) Final report: Sampling-based algorithms for estimating structure in big data. tech. report. Sandia National Laboratory, Livermore
Ng AY, Jordan MI, Weiss Y (2002) On spectral clustering: Analysis and an algorithm. In: NIPS'01 Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. MIT Press, Cambridge, MA. pp 849–856
Noble J, Adams N (2018) Real-time dynamic network anomaly detection. IEEE Intell. Syst. 33:5–18
Nowak K, Marrero CO, Young SJ On the structure of isometrically embeddable metric spaces. arxiv:1808.10509
Price-Williams M, Heard N, Turcotte M (2017) Detecting periodic subsequences in cyber security data. In: 2017 European Intelligence and Security Informatics Conference (EISIC). IEEE. pp 84–90. https://doi.org/10.1109/eisic.2017.40
Ranshous S, Shen S, Koutra D, Harenberg S, Faloutsos C, Samatova NF (2015) Anomaly detection in dynamic networks: a survey. Wiley Interdiscip. Rev. Comput. Stat. 7:223–247
Sanfeliu A, Fu K-S (1983) A distance measure between attributed relational graphs for pattern recognition. In: IEEE Transactions on Systems, Man, and Cybernetics. SMC-13. pp 353–362. https://doi.org/10.1109/tsmc.1983.6313167
Sapienza A, Panisson A, Wu J, Gauvin L, Cattuto C (2015) Anomaly detection in temporal graph data: An iterative tensor decomposition and masking approach. In: International Workshop on Advanced Analytics and Learning on Temporal Data. AALTD 2015, New York
Söderberg B (2002) General formalism for inhomogeneous random graphs. Phys Rev E 66. https://doi.org/10.1103/physreve.66.066121
Sensarma D, Sarma SS (2015) A survey on different graph based anomaly detection techniques. Indian J Sci Technol 8. https://doi.org/10.17485/ijst/2015/v8i1/75197
Siegel S, N.J.C. Jr (1988) Nonparametric Statistics for The Behavioral Sciences. McGraw-Hill Humanities/Social Sciences/Languages, New York
Simard R, L'Ecuyer P (2011) Computing the two-sided kolmogorov-smirnov distribution. J Stat Softw 39. https://doi.org/10.18637/jss.v039.i11
Simpson O, Seshadhri C, McGregor A (2015) Catching the head, tail, and everything in between: A streaming algorithm for the degree distribution. In: 2015 IEEE International Conference on Data Mining. IEEE
Stolman A, Matulef K (2017) HyperHeadTail. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 - ASONAM '17. ACM Press. https://doi.org/10.1145/3110025.3119395
Tong H, Lin C-Y (2011) Non-negative residual matrix factorization with application to graph anomaly detection. In: Proceedings of the 2011 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972818.13
von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17:395–416
Wang J, Paschalidis I. C. (2017) Botnet detection based on anomaly and community detection. IEEE Trans Control Netw Syst 4:392–404
Young IT (1977) Proof without prejudice: use of the kolmogorov-smirnov test for the analysis of histograms from flow systems and other sources. J Histochem Cytochem 25:935–941
Young SJ (2008) Random dot product graphs: a flexible model for complex networks. PhD thesis. Georgia Institute of Technology
Young SJ, Scheinerman E (2008) Directed random dot product graphs. Internet Math 5:91–111
Young SJ, Scheinerman ER (2007) Random dot product graph models for social networks. In: Algorithms and, Models for the Web-Graph. Springer, Berlin Heidelberg. pp 138–149
Zeng Z, Tung AK, Wang J, Feng J, Zhou L (2009) Comparing stars: On approximating graph edit distance. Proc VLDB Endowment 2:25–36

## Publisher's Note