Utilizing Dynamic Time Warping for Pandemic Surveillance: Understanding the Relationship between Google Trends Network Metrics and COVID-19 Incidences

Michael T. Lopez II
Ateneo Social Computing Science Laboratory
Department of Information Systems and Computer Science
Ateneo de Manila University
Quezon City, Philippines
michael.lopez@student.ateneo.edu

Cheska Elise Hung
Ateneo Social Computing Science Laboratory
Department of Information Systems and Computer Science
Ateneo de Manila University
Quezon City, Philippines
cheska.hung@student.ateneo.edu

Maria Regina Justina E. Estuar
Ateneo Social Computing Science Laboratory
Department of Information Systems and Computer Science
Ateneo de Manila University
Quezon City, Philippines
restuar@ateneo.edu

***Abstract*—The premise of network statistics derived from Google Trends data to foresee COVID-19 disease progression is gaining momentum in infodemiology. This approach was applied in Metro Manila, National Capital Region, Philippines. Through dynamic time warping (DTW), the temporal alignment was quantified between network metrics and COVID-19 case trajectories, and systematically explored 320 parameter configurations including two network metrics (network density and clustering coefficient), two data preprocessing methods (Rescaling Daily Data and MSV), multiple thresholds, two correlation window sizes, and Sakoe-Chiba band constraints. Results from the Kruskal-Wallis tests revealed that five of the six parameters significantly influenced alignment quality, with the disease comparison type (active cases vs. confirmed cases) demonstrating the strongest effect. The optimal configuration, which is using the network density statistic with a Rescaling Daily Data transformation, a threshold of 0.8, a 15-day window, and a 50-day radius constraint, achieved a DTW score of 36.30. This indicated substantial temporal alignment with the COVID-19 confirmed cases data. The discoveries demonstrate that network metrics rooted from online search behavior can serve as complementary indicators for epidemic surveillance in urban locations like Metro Manila. This strategy leverages the Philippines' extensive online usage during the pandemic to provide potentially valuable early signals of disease spread, and offers a supplementary tool for public health monitoring in resource-limited situations.**

***Index Terms*—COVID-19, dynamic time warping, network analysis, Google Trends, infodemiology, disease surveillance**

I. Introduction

The Philippines received long-term significant damage from the coronavirus disease (COVID-19) when it was one of the first nations in Southeast Asia that got infected in February 2020 [1]. It is worth looking that those in the medical front lines against the disease in April 2020 were the most affected [2]. The early phase of the pandemic logged 8,212 COVID-19 cases in the entire country, wherein about 24.7% of them were healthcare workers and 35 of those in the medical field perished [3]. At the end of 2020, the Philippines had 451,839 confirmed cases and 8,812 deaths, which makes the case-fatality rate to be 1.95% [4]. Through a heightened sense of urgency, the national government ushered more ramped up activities in contact tracing and testing capacity [4]. It seems this was a positive trait until the Delta variant led to a new wave of cases for the new year that caused the mortality rate to be 33% above normal by the end of 2021 [5].

This pandemic also caused a deep recession in the Philippine economy. The gross domestic product in 2020 shrank by 9.5% due to the strict community lockdowns, spearheading businesses to inactivity [6]. The sharp economic decline, driven by suppressed consumption and investment, was among the worst in the Southeast Asian region for that year [1]. One of those contributing factors were one of the most stringent lockdowns worldwide. The government had its initial classification called the "Enhanced Community Quarantine" imposed on the National Capital Region (also known as Metro Manila), wherein thereafter expanded to the entire Luzon island [1]. Under these circumstances, residents were ordered to stay at home, public transport and non-essential businesses were shut, mass gatherings were banned, and the police and military forces administered the curfews nationwide [2].

The Philippines also had an extensive social media usage during the pandemic [1]. Filipinos turned to online means that augmented with their regular routines for work, education,

and shopping [7]. Another non-pharmaceutical intervention that organically developed during the pandemic were practices of the public looking up information online [8]. There were significant interest at the onset of the pandemic, considering a majority of the information about COVID-19 and its societal implications were novel. Specifically, search engines such as Google has been proven to be a reliable public surveillance tool for virtual crowd behavior [9]. It is worth acknolwedging that the vast majority of infodemiological studies probe selected *individual* keywords on how popular they are on a search engine [10].

The aim of this investigation is to determine whether the network statistics of Google Trends (GT) keyword popularity scores could possibly match the temporal evolution of COVID-19 disease cases through the pattern recognition technique called dynamic time warping (DTW). The additional Sakoe-Chiba radius restriction offers a computationally inexpensive attribute to DTW because it only searches within the temporal space it was given. Another contribution of this study is that it pinpoints a specific geographic location in a country, which is in Metro Manila (an administrative region that has 16 cities and one municipality) with a total population of at least 13.4 million [11]. This deviates from the common scope used in the literature because much of them focused more on entire countries [12]. Through the best of our knowledge, this is the first opportunity to compare retroactive COVID-19 cases during the pandemic and networked online search behavior in the Philippines through a *naive* computation such as DTW.

II. Related Literature

A. Google Trends Mechanism and Reliability

Google Trends (GT) is made available for anyone to find a search term's relative popularity over a certain place and time [13]. Any user across the world may download a specific time-series data of popularity scores based on the keywords or topics that the user queried [14]. These series of popularity scores are called *relative search volume* (RSV) integers that are bounded inclusively within 0 to 100 [15]. There are several attributes that the user may customize to enhance retrieving the GT data: place (this could be filtered from worldwide until down to the city level), query category (such as the parent categories of "Health", "Finance", and "Arts & Entertainment" among others), time period (such that the user may input its preferred start and end dates), and the type of search (limited to "Web Search", "Image Search", "News Search", "YouTube Search", or "Google Shopping") [16].

For the GT query categories, there are an estimated 1,400 possible combinations to choose from because each parent category has at most four hierarchical levels deep [17]. For example, someone may go to "Business & Industrial" $>$ "Computers & Electronics" $>$ "Networking" $>$ "VPN & Remote Access" as one way to filter out the popularity of a search term under that context. GT data is also *anonymized*, which it will then process and aggregate the data in samples instead of the entire population [15]. It is said that Google receives billions of search requests worldwide and at any given moment. Therefore, it is computationally more feasible to calculate a subset of the search inputs at that specific time span [17]. Additionally, the GT data may return at different time intervals depending on the time period the user imposed. These are in hourly, daily, weekly, and monthly temporal configurations [15]. If someone retrieves a time period only from the previous seven days, then the hourly RSV score intervals are returned [16]. There is a maximum of nine months of GT data for someone to download the daily RSV intervals. Then beyond nine months, the weekly data is returned. And lastly, monthly GT data is used for timelines beyond 5.25 years [18].

Conflating these two points, there is a high chance of a potential problem to receive inconsistent GT data. The random sampling algorithm from Google is proprietary, which it is prone to sample variations and may affect the results [18]. Besides this, the calculation of the RSV frequencies regardless of time formats have different data resolutions against one another [17]. Either way, data preprocessing techniques such as normalization are reliable ways to overcome these obstacles and preserve the data [15]. Thus, GT is still a frontier tool to assist the investigation and forecasting of a potential disease outbreak in an area [13]. Google represents 84.08% of the global market share of the search engine industry [16]. This is its biggest advantage because it provides an almost instantaneous real-time statistics to supply initiatives that predict social phenomena [17]. It reliably estimates public perceptions, interests, and behaviors which are helpful for decision-based processes [19].

B. Dynamic Time Warping in Health Analytics

Various initiatives pursued DTW either as the primary tool for disease case surveillance or as an additional tool for a different purpose. Tracking human norovirus, which is the leading cause of acute gastroenteritis in the United States, was the aim of a particular study in Detroit, Michigan [20]. The proponents applied dynamic time warping (DTW) analysis to compare waste water concentrations against Google search trends, and clinical cases whether if such relationship exists. DTW discovered that norovirus levels in wastewater, after being normalized by fecal indicators, produced the most similar trajectories to both the region's positive cases data and internet searches for "norovirus".

In terms of mental health, depression symptom dynamics were focused on a cohort of psychiatric inpatients [21]. DTW was applied to account for possible mismatch in symptom onset or improvement. The computed DTW distance for each symptom pair measured how similar the two symptoms' severity changed across repeated assessments. It unveiled which sets of symptoms tended to peak or subside in unison within and across patients. As a consequence, this approach facilitated more personalized clinical insights about their core depressive symptoms.

On the other hand, there were instances of specific DTW applications on COVID-19 incidences worldwide. The public interest in COVID-19 (gauged through Google Trends data)

were lined up with national-level COVID-19 case and mortality data in six Western nations [22]. DTW was implemented to check the shape-based similarity between the time series of Google search popularity and the time series of COVID-19 deaths and cases, and the government stringency index. They learned that public interest (Google searches) was most closely aligned with mortality. Then, public interest was less closely aligned with raw incident cases, and even lesser against the stringency index.

In Poland, there was an enactment of DTW that focused on the 16 provinces of the country. It analyzed two time series in each voivodeship (province): the cumulative number of infected individuals, and the cumulative number of deaths [23]. Thanks to the DTW distances, it showed that the regions can be grouped based on the shape (and timing) of case and death curves, indicating that local characteristics (such as population density, and proximity to large cities) likely influence how quickly or severely the virus spreads.

Another unique implementation is using DTW as a strategy to cluster COVID-19 cases among 50 nations [24]. Here, the aim of DTW is to identify the set of countries whose daily-cases curves most closely resembled the country to be predicted. By clustering countries in "similar epidemic shapes," it was hypothesized that the relevant patterns would transfer better to the target country's forecast. The ramifications of the "time-series filtering" approach (using DTW to pick the most-similar countries for training) significantly boosted forecasting accuracy. Depending on the machine-learning method, RMSE was reduced by roughly 74% once the DTW-based selection of similar countries were incorporated.

III. Methodology

A. Google Trends Data Retrieval and Preprocessing

The `pytrends` Python (version 3.12) library [25] was imported to automatically scrape the relative search volume (RSV) values of all 15 keywords. These search terms were deliberately chosen so that it satisfies five categories: "cough", "fever", "flu", "headache", and "rashes" for the **English Symptoms** classification. The second category is its **Filipino Symptoms** counterpart, which consists of "lagnat" (fever), "sipon" (runny nose), and "ubo" (cough). For terms on **Face Wearing**, it includes "masks" and "face shield". The **Quarantine** category has only two search terms: "ecq" (Enhanced Community Quarantine) and "quarantine". Lastly, the **New Normal** category monitored regular life implications of the pandemic such as "frontliners", "social distancing", and "new normal". Each search term has a daily RSV time-series data from March 16, 2020 until March 15, 2021 depending on what data preprocessing is used. All the scraping occurred on a 30-day rolling window wherein each comma separated value (CSV) file were segmented beginning on each day within the one-year timeframe until this 30-day window data retrieval is not possible anymore.

Two data preprocessing techniques were used to address the Google Trends' inherent limitation on temporal resolution. First is the *Rescaling Daily Data Method*, which was derived from [26] when they were "stitching" together different sets of weeks to form a cohesive timeline. The method name came to be because it leverages the relationship between daily and weekly GT data to fulfill calibration factors. The weekly GT data, which spans the entire study period in a single query, provides a consistent scaling reference against which the segmented daily data can be calibrated. This plan yielded multiple CSV files per keyword, each containing daily RSV data for a discrete 30-day segment, collectively spanning the entire study period. Weekly RSV data was also gathered directly from GT by requesting the entire period (March 16, 2020, to March 15, 2021) in a single query.

The second data preprocessing strategy adopted was the *Merged Search Volume* (MSV) algorithm [27]. The inventors of the MSV presented a new way to preserve the daily data (or any format) for an extended length of time. Unlike the Rescaling Daily Data Method, which calibrates daily data using weekly references, the MSV transformation *directly addresses* the temporal inconsistency by estimating correction factors between consecutive data segments. This is helpful when seeking to maintain the relative proportionality of daily fluctuations.

B. Conversion of Google Trends Data into Network Statistics

Following the preprocessing of Google Trends data using both transformations, dynamic network graphs were constructed to model the temporal interconnectedness among COVID-19-related search terms. This method extends beyond traditional time-series analysis because it incorporates the complex interdependencies between multiple search queries simultaneously [28]. To capture the evolving relationships between queries, a rolling-window was implemented to construct the time-dependent networks. For each time point t, 15-day and 30-day retrospective correlation windows of preprocessed daily data were considered for all 15 keywords. Using the preprocessed daily RSV data, the distance correlation matrix ρ_t was computed between all pairs of keywords at each time point t using the `dcor` Python package (version 3.12) [29].

To convert these correlation measurements into a network structure, there was a minimum requirement of reaching at least the threshold to establish edges between keyword nodes. The adjacency matrix $A_{i,j,t}$ among search terms i and j at date t was implemented such that if the correlation value is greater than or equal to θ (correlation threshold), then it merits an undirected edge between i and j for that graph [27]. To comprehensively investigate the sensitivity of network metrics to varying levels of connection strength, there was a systematic computation of threshold values: $\theta \in \{0.4, 0.5, 0.6, 0.8\}$. The multi-threshold parameters enable assessments of network dynamics across different stringency levels [28]. Lower thresholds ($\theta = 0.4$) capture weaker associations, resulting in denser networks that may include spurious connections. Higher thresholds ($\theta = 0.8$) impose more restrictive criteria, as it yields sparser networks with only the strongest associations preserved. Intermediate thresholds ($\theta = 0.5, 0.6$) provide

balanced perspectives, with $\theta = 0.5$ serving as the standard threshold in the network analysis literature [30].

To quantify the macroscopic properties of these networks, two complementary network statistics were calculated: *network density* and *clustering coefficient*. The metrics below provide meaningful insights into its structural characteristics. Network density (D_t) represents the proportion of potential connections that are actually realized in the network at time t [31]:

$$D_t = \frac{2E_t}{N(N-1)} \tag{1}$$

where E_t is the number of edges in the network at time t, and $N(N-1)/2$ represents the maximum possible number of edges in an undirected network with N nodes. The values range from 0 (no connections) to 1 (fully connected network). This metric quantifies the overall cohesiveness of health-seeking behavior across the multiple coronavirus pandemic keywords. Periods of high network density suggest synchronized interest in multiple related search terms, potentially indicating collective public response to significant pandemic developments [32]. While network density characterizes overall connectedness, the clustering coefficient (C_t) provides insight into the network's local structure by quantifying the tendency of nodes to form tightly connected groups or clusters [33]. For each node v, there exists $\lambda(v)$, the number of triangles formed with that node, and $\tau(G)$, the total number of triplets across the entire graph. The global clustering coefficient is then calculated as [34]:

$$C_t(G) = \frac{\sum_{v \in V} \lambda(v)}{\tau(G)} = \frac{1}{V} \sum_{v \in V} c(v) \tag{2}$$

The clustering coefficient aggregates these local measures, weighted by each node's connectivity, to provide a network-level assessment of clustering. High values indicate that search terms tend to form cohesive groups, suggesting thematic clusters in health-seeking behavior. Low values suggest more dispersed or random patterns of association between search terms [28]. In the context of the online searches, high clustering may indicate distinct thematic groups of search terms (e.g., symptoms-related queries clustering separately from policy-related terms), while lower clustering despite high density would suggest more uniform associations across different categories of the keywords.

C. Retrieving COVID-19 Daily Confirmed and Active Cases

The Department of Health (DOH) owns a public repository of COVID-19 cases called the data drop, wherein its epidemiology bureau (EB) collects the number of patients who got infected from hospitals nationwide and releases them at noon [35]. This initiative ended in 2023, however the repository is still available at the time of writing [36]. The Metro Manila data was extracted from here such that the `RegionRes` ("region of residence") and `ProvinceRes` ("province of residence") must be equal to `NCR`. The `DateRepConf` ("date of report confirmation") column is the most crucial variable for determining the daily *confirmed* cases because it tells the date in which the patient has been publicly announced as a positive case of the coronavirus [37]. Opposite to the reported infection of an individual is the `DateRepRem` ("date of report of removal") variable, which tells the instance of when was the recovery or death of that patient happened [38]. Both `DateRepConf` and `DateRepRem` are important to calculate the *active* coronavirus cases as the pandemic progressed.

Daily confirmed cases represent the incidence of newly diagnosed COVID-19 cases on each specific date. This metric reflects the rate at which new infections are being detected within the population. In contrast, active cases represent the prevalent burden of COVID-19 at any given time point. It is the accumulated number of confirmed cases minus those who have been removed from the active case pool through recovery or mortality. This metric provides a snapshot of the concurrent disease burden within the community. For each date in this sequence, two key parameters were quantified: the number of new cases confirmed on that date and the number of cases removed (through recovery or death) on that same date. The active case count was then calculated through a cumulative approach, where $A_t = A_{t-1} + C_t - R_t$, such that A_t represents the active cases on day t, C_t denotes new confirmed cases on day t, and R_t signifies removed cases on that same day t. This recurrence relation was initialized at zero and computed iteratively across the entire time series, producing a dynamic representation during the pandemic.

D. Implementation of Dynamic Time Warping

Dynamic time warping (DTW) is commonly used to quantify similarities between time series of different lengths that may exhibit temporal distortions [39]. Consider two time series $X = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^N$ of length N representing daily disease case (active or confirmed) counts and $Y = (y_1, y_2, \ldots, y_M) \in \mathbb{R}^M$ of length M denoting Google Trends data. DTW aims to find an optimal alignment between these sequences by warping the time axis. Formally, DTW identifies a warping path $P = (p_1, p_2, \ldots, p_L)$ where each element $p_l = (i_l, j_l) \in [1:N] \times [1:M]$ pairs index i_l from sequence X with index j_l from sequence Y [40]. A valid warping path must satisfy the following constraints:

- **Boundary condition**: $p_1 = (1,1)$ and $p_L = (N, M)$, ensuring the alignment considers complete sequences.
- **Monotonicity condition**: $i_1 \leq i_2 \leq \ldots \leq i_L$ and $j_1 \leq j_2 \leq \ldots \leq j_L$, preserving the time-ordering of points.
- **Continuity condition**: $p_{l+1} - p_l \in \{(1,0), (0,1), (1,1)\}$ for $l \in [1 : L-1]$, limiting the warping to adjacent cells.

The DTW distance between X and Y is then defined as the minimum cumulative distance over all possible warping paths [23]:

$$\text{DTW}(X, Y) = \min_{P \in \mathcal{P}} \sum_{l=1}^{L} d(x_{i_l}, y_{j_l}) \tag{3}$$

where $\mathcal{P}$ is the set of all valid warping paths and $d(x_{i_l}, y_{j_l})$ represents the distance between elements x_{i_l} and y_{j_l}. Given the inherent differences in scale between these series, the disease case data requires normalization prior to comparison. The min-max normalization was applied to the disease case time series as follows:

$$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{4}$$

The Google Trends data is maintained in its original form as it is because the network statistics values are already between 0 and 1. It is known that DTW employs dynamic programming to efficiently compute the optimal warping path [39]. The approach constructs an accumulated cost matrix $D \in \mathbb{R}^{N \times M}$ where each element $D(i, j)$ represents the minimum cumulative distance for aligning subsequences $X_{\text{norm}}(1 : i)$ and $Y(1 : j)$. The local cost matrix $C \in \mathbb{R}^{N \times M}$ was first computed:

$$C(i, j) = |x_{\text{norm},i} - y_j|, \quad i \in [1 : N], j \in [1 : M] \tag{5}$$

Then, the accumulated cost matrix is populated recursively:

$$D(i, j) = C(i, j) + \min \begin{cases} D(i-1, j) \\ D(i, j-1) \\ D(i-1, j-1) \end{cases} \tag{6}$$

with base case $D(1, 1) = C(1, 1)$. Typically, unrestricted DTW is oftenly practiced. However, applying this characteristic produces alignments where a single point in one series maps to a large subsection of the other. To prevent this and improve computational efficiency, the Sakoe-Chiba band constraint was considered [41], which restricts the warping path to a band along the main diagonal $|i - j| \leq r$, wherein r is the radius parameter defining the maximum allowable temporal deviation in terms of the number of days. This constraint modifies the accumulated cost matrix calculation:

$$D(i, j) = \begin{cases} C(i, j) + \min\{D(i-1, j), \\ \quad D(i, j-1), D(i-1, j-1)\}, & \text{if } |i - j| \leq r \\ \infty, & \text{if } |i - j| > r \end{cases} \tag{7}$$

For our analysis of the relationship between disease cases and Google network statistics behavior, multiple radius values were evaluated: $r \in \{7, 15, 20, 30, 50\}$ days. The constraint creates a corridor of width $2r + 1$ along the matrix diagonal, significantly reducing the search space from $O(NM)$ to $O(Nr)$ wherein $r < M$. After computing the accumulated cost matrix, the optimal warping path is retrieved through backtracking. The resulting path $P^* = (p_1, p_2, \ldots, p_L)$ provides the optimal element-wise mapping between disease case patterns and networked search behavior indicators. This implementation uses Python with the `numpy` package for matrix operations. The Sakoe-Chiba constraint is applied via a Boolean mask matrix:

$$M(i, j) = \begin{cases} \texttt{False}, & \text{if } |i - j| \leq r \\ \texttt{True}, & \text{if } |i - j| > r \end{cases} \tag{8}$$

where cells marked `True` are excluded from the path calculation. For each disease-search term pair, DTW distances were determined across all radius values to determine the optimal temporal constraint. Whatever the result of the distance is, this serves as the similarity metric. It tells that lower values have stronger correspondence between disease patterns and search behavior. Provided below is Algorithm 1 that summarizes how the DTW was evaluated between the two time-series.

Algorithm 1 DTW with Sakoe-Chiba Band

Require: Series $X \in \mathbb{R}^N$, $Y \in \mathbb{R}^M$, radius r
Ensure: DTW distance $D(N, M)$ and optimal path P^*
1: $X_{\text{norm}} \leftarrow \frac{X - \min(X)}{\max(X) - \min(X)}$
2: Initialize cost matrix $C(i, j) \leftarrow |x_{\text{norm},i} - y_j|$ for all i, j
3: Initialize D with ∞; $D(1, 1) \leftarrow C(1, 1)$
4: **for** $i = 1$ to N, $j = 1$ to M where $|i - j| \leq r$ **do**
5: **if** $(i, j) \neq (1, 1)$ **then**
6: $D(i, j) \leftarrow C(i, j) + \min \begin{cases} D(i-1, j) & \text{if } i > 1 \\ D(i, j-1) & \text{if } j > 1 \\ D(i-1, j-1) & \text{if } i, j > 1 \end{cases}$
7: **end if**
8: **end for**
9: Backtrack from (N, M) to $(1, 1)$ to build P^*
10: **return** $D(N, M)$, P^*

IV. Results and Discussion

The network metrics were constructed from Google Trends data using both the Rescaling Daily Data Method and Merged Search Volume (MSV) algorithm transformations. We systematically explored 320 distinct parameter configurations to identify optimal settings for epidemic surveillance through network analysis. The parameter space encompassed two network metrics (network density and clustering coefficient), two data source transformations (Rescaling Daily Data and MSV), four threshold values (0.4, 0.5, 0.6, and 0.8), two correlation window sizes (15 and 30 days), two case comparison types (Confirmed Cases and Active Cases), and five Sakoe-Chiba band radius values (7, 15, 20, 30, and 50 days). DTW scores in the dataset ranged from 36.30 to 101.67. The mean DTW score across all configurations was 62.30. This mean serves as the baseline for assessing relative performance.

A. Statistical Significance of the Parameters

To systematically evaluate the impact of each parameter on DTW performance, the non-parametric Kruskal-Wallis tests were conducted, which serves as an alternative to one-way ANOVA [42]. Table I presents the results, revealing that five of the six parameters significantly influenced alignment quality ($p < 0.05$).

The most influential parameter was the disease case comparison dimension ($H = 154.80$, $p = 1.55 \times 10^{-35}$), with

TABLE I
KRUSKAL-WALLIS TEST RESULTS FOR PARAMETER SIGNIFICANCE

Variable	H-statistic	p-value	Significant?
Network Statistic Type	26.44	2.72e-7	Yes
Data Preprocessing Technique	4.10	4.30e-2	Yes
Threshold	27.71	4.17e-6	Yes
Correlation Window	1.66	1.98e-1	No
Disease Case Comparison	154.80	1.55e-35	Yes
Sakoe-Chiba Radius	20.97	3.21e-4	Yes

confirmed cases yielding substantially better alignment than active cases (average DTW scores of 51.08 and 73.53, respectively). This finding suggests that network metrics derived from Google Trends activity in Metro Manila correlate more closely with new case identification than with the prevalence of active infections. This pattern may reflect public health-seeking behavior in response to official case announcements, which typically emphasize daily new cases rather than active case counts in health-seeking behavior.

Simultaneously, the threshold values parameter significantly impacted performance ($H = 27.71$, $p = 4.17 \times 10^{-6}$), with lower thresholds of 0.5 (avg. DTW = 58.30) and 0.4 (avg. DTW = 58.48) generally outperforming more restrictive thresholds of 0.8 (avg. DTW = 71.11). This indicates that moderate thresholds for establishing connections between search terms in our network model captured more meaningful dynamics during the pandemic in Metro Manila. Too high thresholds may have eliminated important but moderate correlations between search terms, while moderate thresholds preserved these relationships.

The network statistic parameter emerged as another significant factor ($H = 26.44$, $p = 2.72 \times 10^{-7}$), with the network density (avg. DTW = 57.07) consistently outperforming against the clustering coefficient (avg. DTW = 67.54). This suggests that the overall density of the search query network was more predictive of epidemic progression in Metro Manila than the tendency of search terms to form tightly connected clusters. During a pandemic where information-seeking behavior spans multiple related topics, the connectivity between topics appears to be more informative than the clustering of specific keyword nodes.

The radius parameter for the Sakoe-Chiba band exhibited significant impact ($H = 20.97$, $p = 3.21 \times 10^{-4}$), with larger radii generally yielding better performance. The average DTW score decreased monotonically as radius increased: 67.75 for the 7-day limitation, 64.11 for the 15-day restriction, 62.62 for the 20-day attribute, 60.14 for the 30-day, and 56.90 for the 50-day imposed parameter. This trend indicates that broader temporal alignment constraints enhance the correspondence between network metrics and epidemic curves. It also likely accommodates more the lag between disease incidence and related online search behavior in Metro Manila's context. The optimal performance at the Sakoe-Chiba restriction of 50 days suggest that public health-seeking behavior has a complex temporal relationship with case trajectories that spans several weeks.

The method on what data preprocessing transformation used also significantly influenced performance ($H = 4.10$, $p = 4.30 \times 10^{-2}$), with the Rescaling Daily Data Method (avg. DTW = 60.48) providing better alignment than MSV (avg. DTW = 64.13). This suggests that the Rescaling Daily Data Method, which leverages the relationship between daily and weekly Google Trends data, better preserves the temporal dynamics relevant to epidemic surveillance than the MSV algorithm. On the other hand, the correlation window size parameter interestingly did not significantly affect performance ($H = 1.66$, $p = 1.98 \times 10^{-1}$), indicating that both 15-day and 30-day windows adequately capture the relevant temporal dynamics for alignment. This robustness to window size variation is advantageous for practical implementation so that it offers some flexibility in temporal aggregation.

B. Optimal Parameter Configurations

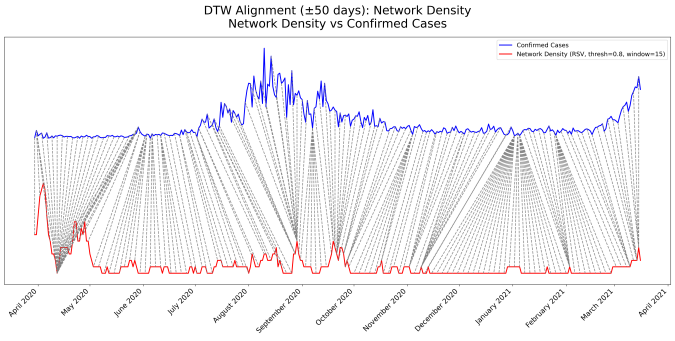

Fig. 1. Temporal Alignment Between Network Density Values (Rescaling Daily Data Method, Threshold = 0.8, Correlation Window = 15 days, Sakoe-Chiba Radius = 50 days) with Confirmed COVID-19 Cases in Metro Manila from March 2020 to March 2021.

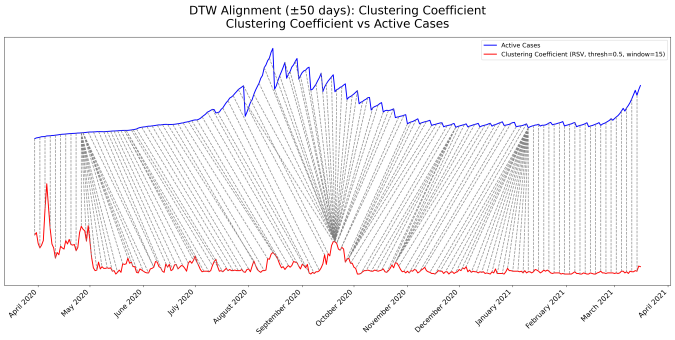

Fig. 2. Temporal Alignment Between Clustering Coefficient Values (Rescaling Daily Data Method, Threshold = 0.5, Correlation Window = 15 days, Sakoe-Chiba Radius 50 day) with Confirmed COVID-19 Cases in Metro Manila from March 2020 to March 2021.

Table II presents the optimal parameter configurations for each combination of metric type and comparison type. Across all combinations, several patterns emerge: larger radii (particularly 50) consistently yield better alignment, the Rescaling Daily Data outperforms MSV as a data transformation method, and optimal thresholds vary based on the specific metric-comparison combination.

TABLE II
OPTIMAL PARAMETER CONFIGURATIONS

Configuration	Data Preprocess	Thr	Win	Rad	DTW
ND - Confirmed	Rescaling Daily	0.8	15	50	36.30
ND - Active	Rescaling Daily	0.6	15	50	42.68
CC - Confirmed	Rescaling Daily	0.5	15	50	36.46
CC - Active	Rescaling Daily	0.4	15	50	51.88

ND: Network Density, CC: Clustering Coefficient
Src: Source, Thr: Threshold, Win: Correlation Window, Rad: Radius

The best overall performance (DTW = 36.30) was achieved with network density, using the Rescaling Daily Data transformation, a threshold of 0.8, a 15-day correlation window, and a radius of 50 days when aligning with confirmed cases. This configuration represents the most promising approach for operationalizing network metrics as epidemic surveillance tools in Metro Manila. Figure 1 illustrates the temporal alignment between this network metric configuration and the confirmed COVID-19 case trajectory.

For active cases, which generally proved more challenging to align with network metrics, the optimal configuration utilized network density from the Rescaling Daily Data Preprocessing Method with a lower threshold of 0.6, 15-day correlation window, and a 50-day Sakoe-Chiba radius (DTW = 42.68). The need for a lower threshold when aligning with active cases may reflect the differing nature of online searches. Meanwhile, clustering coefficient metrics exhibited their best performance when utilizing the same Rescaling preprocessing method with moderate thresholds (0.5 for confirmed cases, 0.4 for active cases), consistently favoring a 15-day correlation window and 50-day Sakoe-Chiba radius. The optimal DTW scores for the clustering coefficient statistic (36.46 for confirmed cases, 51.88 for active cases) were comparable to or slightly worse than those achieved with the network density. This again reinforces the preference for density-based metrics in this application.

C. Implications for Epidemic Surveillance in the Philippine Context

The results demonstrate that Google Trends-based network metrics, when optimally configured, can achieve substantial temporal alignment with COVID-19 case trajectories in Metro Manila. The best configurations achieved DTW scores in the 36 to 52 range, representing significantly better alignment than would be expected by chance. It suggests that network metrics derived from online search behavior data can serve as complementary indicators for traditional epidemic surveillance systems in the Philippine setting. These results are particularly relevant given the Philippines' extensive digital presence during the pandemic [1]. As Filipinos increasingly turned to online means to support their regular routines for health-seeking actions during strict community quarantine measures [7], the digital traces left became increasingly valuable for public health monitoring.

The superior performance of the network density metrics emphasizes the value of overall connectivity patterns in tracking epidemic progression. As information about disease outbreaks diffuses throughout online, the density of connections appears to connect strongly with case trajectories. Thus, it has the potential of serving as a leading indicator for epidemic waves. This insight implies that changes in online health-seeking patterns may often respectively precede official case reporting.

V. CONCLUSION

This comprehensive parameter exploration for aligning Google Trends search popularity network metrics with COVID-19 case trajectories in Metro Manila provides key insights for epidemic surveillance in a local context. First, the network density consistently outperforms the clustering coefficient for this application, though both metrics can achieve strong alignment when optimally configured. Second, confirmed cases generally align better with network metrics than active cases, suggesting that online health-seeking activity more closely tracks new case identification than ongoing prevalence. Third, larger temporal alignment radii (such as 50 days) consistently improve mapping quality. It has repercussions on how anticipating future disease cases could take a significant number of days prior. In return, government health agencies and officials may benefit on the proactivity of detecting cases way ahead of time.

The significance of five out of six parameters in the analysis underscores the importance of careful parameter selection when implementing network-based surveillances. The non-significance of window size provides valuable flexibility for practical implementation, allowing system designers to balance temporal resolution against computational efficiency without compromising alignment quality. Therefore, our results demonstrate that a network metric approach towards Google Trends data, when appropriately configured, can serve as valuable complementary indicators for epidemic surveillance in the Metro Manila region. By capturing the complex patterns of public information-seeking behavior during a health crisis, these metrics provide insights beyond traditional surveillance data, which has the potential to discover early warning capabilities and appropriately strategize on its communication strategies for future outbreaks.

ACKNOWLEDGMENTS

The authors forward its gratitude to the Ateneo Social Computing Science Laboratory and its parent organization, the Ateneo Center for Computing Competency and Research, for the opportunity and platform of spearheading this study. Much of the appreciations are also extended to Christian Pulmano and Atty. Kennedy Espina for all their suggestions and recommendations shared towards the success of this endeavor.

REFERENCES

[1] A. M. Amit, V. C. Pepito, and M. Dayrit, "Early response to COVID-19 in the Philippines," *West. Pac. Surveill. Response J.*, vol. 12, no. 1, pp. 56–60, Mar. 2021.

[2] M. K. B. De Vero, I. E. A. Gozum, and A. M. Melad, "Prevention and planning as important factors in ensuring public health in the Philippines during the COVID-19 pandemic," *J. Public Health*, vol. 43, no. 2, Jun. 2021.

[3] E. W. A. Leyva, J. I. D. Soberano, J. T. Paguio, K. L. L. Siongco, E. F. R. Sumile, and S. R. Bonito, "Pandemic Impact, Support Received, and Policies for Health Worker Retention: An Environmental Scan," *Acta Med. Philipp.*, vol. 58, no. 12, Jul. 2024.

[4] World Health Organization, "Philippines: Coronavirus Disease 2019 (COVID-19) Situation Report #65," World Health Organization, Manila, Tech. Rep., Dec. 2020.

[5] J. R. Migriño and M. R. Bernardo-Lazaro, "Using an online calculator to describe excess mortality in the Philippines during the COVID-19 pandemic," *Western Pac. Surveill. Response J.*, vol. 14, no. 1, pp. 1–11, Mar. 2023.

[6] AlterContacts, "Lockdown Economy Philippines." [Online]. Available: https://www.altercontacts.org/report/lockdown-economy-philippines

[7] J. Caparino, C. Magahis, and J. Santua, "Filipinos' reliance on internet at an all-time high," https://manilastandard.net/?p=357427, Jun. 2021.

[8] A. Galido, J. J. Ecleo, A. Husnayain, and E. Chia-Yu Su, "Exploring online search behavior for COVID-19 preventive measures: The Philippine case," *PLoS One*, vol. 16, no. 4, Apr. 2021.

[9] A. Sofi-Mahmudi, E. Shamsoddin, P. Ghasemi, M. Nasser, and B. Mesgarpour, "Assessing Google Searches for Toothache during COVID-19 Lockdowns," *Med. J. Islam. Repub. Iran*, vol. 37, no. 1, pp. 282–293, Mar. 2023.

[10] B. A. Zayed, A. M. Talaia, M. A. Gaaboobah, S. M. Amer, and F. R. Mansour, "Google Trends as a predictive tool in the era of COVID-19: A scoping review," *Postgrad. Med. J.*, vol. 99, no. 1175, pp. 962–975, Sep. 2023.

[11] Philippine Statistics Authority, "Highlights of the National Capital Region (NCR) Population 2020 Census of Population and Housing (2020 CPH)," https://psa.gov.ph/statistics/population-and-housing/node/165009, Aug. 2021.

[12] T. Saegner and D. Austys, "Forecasting and Surveillance of COVID-19 Spread Using Google Trends: Literature Review," *Int. J. Environ. Res. Public Health*, vol. 19, no. 19, Jan. 2022.

[13] A. Rovetta, "Google trends in infodemiology: Methodological steps to avoid irreproducible results and invalid conclusions," *Int. J. Med. Inform.*, vol. 190, Oct. 2024.

[14] P. Behnen, R. Kessler, F. Kruse, J. M. Gómez, J. Schoenmakers, and S. Zerr, "Experimental Evaluation of Scale, and Patterns of Systematic Inconsistencies in Google Trends Data," in *Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020)*, ser. Communications in Computer and Information Science, I. Koprinska, M. Kamp, A. Appice, C. Loglisci, L. Antonie, A. Zimmermann, R. Guidotti, Ö. Özgöbek, R. P. Ribeiro, R. Gavaldà, J. Gama, L. Adilova, Y. Krishnamurthy, P. M. Ferreira, D. Malerba, I. Medeiros, M. Ceci, G. Manco, E. Masciari, Z. W. Ras, P. Christen, E. Ntoutsi, E. Schubert, A. Zimek, A. Monreale, P. Biecek, S. Rinzivillo, B. Kille, A. Lommatzsch, and J. A. Gulla, Eds. Ghent, Belgium: Springer International Publishing, 2020, vol. 1323, pp. 374–384.

[15] E. Cebrián and J. Domenech, "Addressing Google Trends inconsistencies," *Technol. Forecast. Soc. Change*, vol. 202, May 2024.

[16] R. Alibudbud, "Google Trends for health research: Its advantages, application, methodological considerations, and limitations in psychiatric and mental health infodemiology," *Front. Big Data*, vol. 6, 2023.

[17] I. Lolić, M. Matošec, and P. Sorić, "DIY google trends indicators in social sciences: A methodological note," *Technol. Soc.*, vol. 77, Jun. 2024.

[18] V. Z. Eichenauer, R. Indergand, I. Z. Martínez, and C. Sax, "Obtaining consistent time series from Google Trends," *Econ. Inq.*, vol. 60, no. 2, pp. 694–705, 2022.

[19] M. C. Medeiros and H. F. Pires, "The Proper Use of Google Trends in Forecasting Models," Apr. 2021.

[20] H. P. Guzman, L. Zhao, M. J. Swain, R. A. Faust, and I. Xagoraraki, "Tracking Norovirus in Tri-County Detroit, MI, Using Wastewater Testing, Syndromic Data, and Online Publicly Available Sources," *ACS EST Water*, vol. 4, no. 11, pp. 4990–5001, Nov. 2024.

[21] K. Hebbrecht, M. Stuivenga, T. Birkenhäger, M. Morrens, E. I. Fried, B. Sabbe, and E. J. Giltay, "Understanding personalized dynamics to inform precision medicine: A dynamic time warp analysis of 255 depressed inpatients," *BMC Med.*, vol. 18, no. 1, Dec. 2020.

[22] P. D. Ziakas and E. Mylonakis, "Public interest trends for Covid-19 and alignment with the disease trajectory: A time-series analysis of national-level data," *PLoS Digit. Health*, vol. 2, no. 6, Jun. 2023.

[23] J. Landmesser, "The use of the dynamic time warping (DTW) method to describe the COVID-19 dynamics in Poland," *Oecon. Copernic.*, vol. 12, no. 3, pp. 539–556, Sep. 2021.

[24] L. Miralles-Pechuán, A. Kumar, and A. L. Suárez-Cetrulo, "Forecasting COVID-19 cases using dynamic time warping and incremental machine learning methods," *Expert Syst.*, vol. 40, no. 6, Jul. 2023.

[25] "Pytrends: Pseudo API for Google Trends." [Online]. Available: https://pypi.org/project/pytrends/

[26] A. Brodeur, A. E. Clark, S. Fleche, and N. Powdthavee, "COVID-19, lockdowns and well-being: Evidence from Google Trends," *J. Public Econ.*, vol. 193, Jan. 2021.

[27] A. M. Y. Chu, A. C. Y. Chong, N. H. T. Lai, A. Tiwari, and M. K. P. So, "Enhancing the Predictive Power of Google Trends Data Through Network Analysis: Infodemiology Study of COVID-19," *JMIR Public Health Surveill.*, vol. 9, Sep. 2023.

[28] M. K. P. So, A. M. Y. Chu, A. Tiwari, and J. N. L. Chan, "On topological properties of COVID-19: Predicting and assessing pandemic risk with network statistics," *Sci. Rep.*, vol. 11, no. 1, Mar. 2021.

[29] C. Ramos-Carreño and J. L. Torrecilla, "Dcor: Distance correlation and energy statistics in Python," *SoftwareX*, vol. 22, May 2023.

[30] M. K. P. So, A. Tiwari, A. M. Y. Chu, J. T. Y. Tsang, and J. N. L. Chan, "Visualizing COVID-19 pandemic risk through network connectedness," *Int. J. Infect.*, vol. 96, pp. 558–561, Jul. 2020.

[31] H. D. Bedru, S. Yu, X. Xiao, D. Zhang, L. Wan, H. Guo, and F. Xia, "Big networks: A survey," *Comput. Sci. Rev.*, vol. 37, Aug. 2020.

[32] A. M. Y. Chu, A. Tiwari, and M. K. P. So, "Detecting early signals of COVID-19 global pandemic from network density," *J. Travel Med.*, vol. 27, no. 5, Aug. 2020.

[33] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, Jun. 1998.

[34] G. Chalancon, K. Kruse, and M. M. Babu, "Clustering Coefficient," in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds. New York, NY: Springer, 2013, pp. 422–424.

[35] Department of Health of the Republic of the Philippines, "Privacy and Confidentiality Statement of the COVID-19 Data Drop," https://drive.google.com/drive/folders/1ZPPcVU4M7T-dtRyUceb0pMAd8ickYf8o?usp=sharing.

[36] ——, "COVID-19 Case Tracker," https://doh.gov.ph/diseases/covid-19/covid-19-case-tracker/.

[37] J. Galarosa and S. Peniano, "How good (or bad) is the DOH reporting COVID-19 data?" https://medium.com/up-scientia/how-good-or-bad-is-the-doh-reporting-covid-19-data-46f0ae00cd07, Sep. 2020.

[38] UP Media and Public Relations Office, "Prevailing Data Issues in the Time of COVID-19 and the Need for Open Data," May 2020.

[39] P. Senin, "Dynamic Time Warping Algorithm Review," 2008.

[40] M. Muller, *Dynamic Time Warping*. Berlin, Heidelberg, Germany: Springer, 2007, pp. 69–84.

[41] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Signal Process.*, vol. 26, no. 1, pp. 43–49, Feb. 1978.

[42] J. I. E. Hoffman, "Chapter 25 - Analysis of Variance. I. One-Way," in *Basic Biostatistics for Medical and Biomedical Practitioners (Second Edition)*, J. I. E. Hoffman, Ed. Academic Press, Jan. 2019, pp. 391–417.