IP Reputation Scoring with Geo-Contextual Feature Augmentation

Henanksha Sainani[1], Josephine M. Namayanja[2], Guneeti Sharma[2], Vasundhara Misal[1], Vandana P. Janeja[1*]
[1]University of Maryland, Baltimore County
[2]University of Massachusetts , Boston
*contact: vjaneja@umbc.edu
Preprint version

1 Introduction

Tobler's [1] First Law of Geography states that everything is related to everything else but nearby objects are more related than distant objects. Thus, geography provides a way to characterize the space where any given phenomena occurs. In this paper, we hypothesize that by combining traditional network attributes with geospatial knowledge, we can derive a much more robust reputation scoring model for IP addresses. One example where this is particularly useful is when there is only header based network data due to encryption of traffic.
We propose a novel approach that combines traditional domain knowledge with additional contextual knowledge in the form of geospatial attributes from multiple independent sources, leading to a much more robust anomaly detection model for the cybersecurity domain. We argue that geo-contextual features are useful when there is limited session data such that a blacklisted IP address operates in a stealth mechanism. While network encryption is the most effective way to attain data security and privacy, attacks against encrypted protocols are increasing significantly. Decryption of such network information not only minimizes security and privacy, it is also extremely expensive and thus not a practical solution for cyberattack detection and prevention. Attacks have exploited this by seriously undermining the ability of traditional Network Intrusion Detection Systems (NIDS), which are heavily dependent on inspecting the network payload and thus are not able to inspect encrypted network sessions effectively. In that regard, a network analyst has to largely depend on unencrypted observables such as an IP address. Additionally, when the session is encrypted, there is little network information available to decide if the IP address is bad. On the other hand, Host Intrusion Detection Systems (HIDS) are host-centric to monitor and audit logs, thus making their implementation costlier as well as operating system specific. More so, the efficiency of HIDS is constrained in real-time, hence unable to circumvent malicious hosts on a network.

With sophisticated attacks increasing the vulnerability of networks, efforts are directed to the utilization of blacklists, a list of IP addresses whose reputation 'describes' to be malicious. However, dependency on blacklists is a constrained approach for intrusion detection and prevention because maintaining updated lists is a challenge. More so, it is argued that blacklists are unreliable due to the sophisticated and relentless nature of attackers to disguise attack behavior by modifying features associated to known malicious IP addresses. More than often dealing with an IP address whose reputation is unknown, becomes challenging. Unlike signature-based IDS that rely on pre-existing knowledge of previous attack signatures, we propose a proactive approach to detect malicious IP addresses based on an anomaly-based IP reputation scoring model. This is because anomaly-based approaches are more effective in detecting new attacks characterized by deviations in the network behavior caused by a cyberattack.

Therefore, we propose supplementing the little-known network information about the IP address with contextual data, specifically augmenting geo-contextual knowledge to traditional network knowledge to create geo-augmented network data. We argue that this provides the meta-knowledge necessary to make an IDS more robust and expand a network analyst's ability to make better decisions. Our proposed model utilizes combined similarity to detect **Geo-Augmented Network Similarity** (**GeoNet**) in IP behavior. Furthermore, our model applies **Reputation Scoring** for IP addresses in real time based on the derived geo-augmented network knowledge

(**GeoNet-RS**). Our scoring model will carry out the following functions: i) take an IP address, ii) extract IP-related features and the geo-location of the IP address, iii) augment IP-related data with geo-contextual data, iv) generate similarity score for an IP address based on baseline model and v) generate a reputation score for an IP address. Network analysts can use this reputation score as supplementary information to decide if this IP address needs to be blacklisted. In general, in an attack scenario when the analyst is deluged with data it is essential to provide them with augmented information, which can help them better evaluate active threats.

Around the world, Information and Communication Technology (ICT) is the driver of various social, economic, demographic and governmental changes. With these developments in play, cybersecurity threats can no longer be viewed in isolation and solely with traditional cybersecurity features. CrowdStrike's 2015 Global Threat Report surfaces that today's global threats are being led by more geopolitical and economic events around the world [2]. The key driver for a global cybersecurity activity has now shifted from disparate activities carried out by individuals, groups and criminal gangs pursuing short-term gains to skilled attackers inspired by strategic global issues. In 2013, Microsoft explored the various technical and non-technical factors that contribute to cybersecurity [3] to show that the prevalence of malware correlates with a variety of economic, social development, ICT policy and technological factors of a country. The US Army Training and Doctrine Command (TRADOC) states that the virtual environment consists of four different types of network layers - data, device, network, and geographic layers [4]. Each of these four layers has nodes, which can be mapped individually to a geographic location. Thus, geography provides a common ground that serves as the foundation required to establish shared situational awareness amongst all these four layers. We are convinced that geographical context matters when assessing the cybersecurity landscape. A much more robust IP reputation scoring model would be developed from boosting traditional network knowledge with geo-contextual information.

1.1 Motivation Example:

Consider a motivating example illustrated in Figure 1 that explains the need for geo-augmentation in the cybersecurity domain. Suppose there are two distinct network sessions involving two IP addresses x and y originating in different parts of Country A from the Eastern European region. If the network administrator has to make a decision whether to blacklist these IP addresses solely on the basis of limited network information available, he/she will have a narrow perspective.

Figure 1: Geo-Contextual Scenario

On the other hand, if the network administrator gets additional knowledge that many neighboring countries in some regions in Eastern Europe host a higher proportion of malware attacks [5] worldwide because of a combination of geo-contextual factors such as reasonable computing infrastructure combined with rampant corruption that allows these malware attackers to flourish, this additional situational awareness expands the network analyst's ability to make better decisions. However, this information is not enough because not all IPs from Country A are malicious IPs. Now the network analyst is provided with additional information that the Internet Service Provider (ISP) of IP address x, who is within Country A is in fact a malicious ISP that has hosted a major portion of cybercrime worldwide. This additional network information along with geo-contextual information enables the network analyst to confidently blacklist IP address x and not necessarily blacklist IP address y. Therefore, the geo-augmentation of the network information allowed the network administrator to look at these two IP addresses with a much different perspective.

This example makes a strong case of augmenting geo-contextual knowledge in the cybersecurity domain. Given that certain locations such as rich countries or big cities are likely to be targets of malware activity compared to poor countries or small cities, by observing just a single geographical feature such as country or city, we lose the ability to explain why certain locations pose as better targets or hosts for malware activity. Therefore, we examine multiple geo-contextual features and argue that overall such a holistic view can explain a key part of the cyber environment and should be a key ingredient of predictive cybersecurity modeling.

With an aim to group similarly behaving IP addresses with respect to their network behavior and geo-contextual information group, this study proposes clustering to generate a baseline model that is used by the reputation scoring model for comparing an IP address. The assumption is that blacklisted IPs exhibit common malicious behavior and can therefore be grouped together into clusters. In real-world applications however, network data features are rarely homogeneous, often containing both continuous and categorical features. It becomes increasingly difficult to find meaningful and well-formed clusters when data has heterogeneous features. More so, the inclusion of meta-knowledge such as geo-contextual features amplify complexity of clustering using heterogeneous features which limits traditional clustering algorithms. We bridge this gap by utilizing an improved clustering algorithm called unified clustering that combines categorical and continuous features to generate well-formed clusters [6]. Our novel reputation scoring approach takes an IP address, compares it against a clustering-based baseline model and produces a trustworthiness score through probabilistic scoring.

1.2 Scenarios with IP Spoofing:

One potential challenge to our proposed reputation model is the vulnerability of an IP address to be spoofed such that an attacker can illicitly impersonate another host by using a forged IP address. We envision our proposed model to work as a supplementary model that bolsters a network analyst's existing restricted approach due to network encryption. The proposed IP reputation model should not be used for network traffic with high probability of being spoofed, for example in UDP traffic. Even in such cases, our model has merit unlike signature-based methods that are dependent on known signatures. We argue that our proposed model is not restricted to network information but leverages the utilization of geo-contextual information as well, potentially enhancing detection accuracy. Therefore, our model is based on assigning a reputation score to an IP address based on similarly behaving IP addresses which includes spoofed IP addresses as well.

Contributions: This study has implications for anomaly detection for cybersecurity applications, especially when there is limited information about the network session or lack of historical data for the network features. Thus, this study's contributions are three-fold;

a) First, we show that our approach of augmenting geo-contextual features to network features produces an improved and robust baseline model for evaluating an IP address in real time. Our approach is not restricted to network information, but leverages information from multiple geo-contextual viewpoints, thus providing a holistic view and potentially enhancing detection accuracy.

b) Second, our baseline model utilizes unified clustering on heterogeneous features to detect similarity in network behavior. We show that using unified clustering outperforms traditional clustering techniques, particularly k-means clustering.

c) Third, we propose a novel reputation scoring approach that compares an IP address in real time, against the proposed baseline model and produces a reputation score through probabilistic scoring. Our approach has implications in the domain of anomaly detection for encrypted sessions when little information is

available to a network analyst. Therefore, our scoring model serves the twin goals of data privacy preservation and anomaly detection.

The rest of the paper is organized as follows: in Section 2 we discuss related work, in Section 3 we discuss the system architecture and methodology, in Section 4, we present experimental results and key findings, in Section 5, we provide a discussion and implications of our study, and lastly in Section 6, we present our conclusions and future work.

2 Related Work

In this section, we discuss the prior work pertaining to i) clustering heterogeneous features, ii) the utilization of geographical information in cybersecurity and ii) reputation scoring approaches for cyber security.

2.1 Clustering Heterogeneous Features

Clustering methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria. Existing clustering methods such as k-means and its variants [7, 8] do not efficiently handle large heterogeneous datasets that contain mixed attributes. Such traditional clustering algorithms employ distance measures to determine similarity between data objects, which makes them more suitable for numerical (continuous) data. As a result, they do not work well for categorical data due to its discrete nature. Some traditional approaches propose to convert categorical data into numeric values [9]. However, this does not necessarily produce meaningful results in the case where categorical domains are not ordered. In Ralambondrainy's [10] study, he proposes to convert multiple category attributes into binary attributes, that is, using 0 and 1 to represent either a category absent or present. In turn, the binary attributes are treated as numeric for which k-means clustering is applied. From the perspective of clustering only categorical attributes, Huang and Ng [11] proposed the k-modes algorithm, which extends k-means to categorical domains whilst preserving the efficiency of the k-means algorithm. Huang & Ng [11] argues that the biggest advantage of k-modes is that it is scalable to very large data sets.

In an attempt to cluster mixed datasets, other studies apply a split-based approach where the mixed dataset in divided into two datasets; one containing continuous attributes and another containing categorical attributes. Here a specific clustering technique suitable to each partition is applied and the intermediate results are integrated [12, 13, 14]. In such cases, the clustering outcomes will depend on the individual clustering techniques applied to different datasets.

On the other hand, a number of studies have proposed efficient approaches for clustering mixed data together such as [15, 16, 17, 18, 19]. Specifically, Huang [15] proposed the k-prototype algorithm to address the problem of clustering large mixed datasets. In their study, they combine k-means and k-modes by applying it to the continuous and categorical attributes respectively. However, the effects of large high dimensional data on clustering outcomes is still an open challenge. Thus, the quality of clusters formed is still questionable. Overall, while previous studies have proposed various approaches for handling mixed data clustering, few methods attempt to cluster heterogeneous mixed data together. Additionally, the similarity measures proposed especially for categorical data may not truly represent the inherent nature of the datasets involved.

2.2 Geo-contextual Perspective on Cybersecurity

Several studies have examined the role of geographical contexts in cyber threat assessment [5, 20, 21]. For example, Mezzour [5] points to the role of technical, political, economic and social factors in a country characterized by computing and monetary resources, cyber security expertise, wealth of residents, computer piracy, international relations, to mention a few. [5, 21] assert that rich countries are likely to be targets of malware activity compared to poor countries. On the other hand, significant computing resources coupled with corruption lend great support to cyber-criminal activity. While [5] points to specific reasons for international variation in initiating and directing attacks, [20] assesses the interdependence of cyberattacks between countries. The approach described in [20] takes into consideration that geographical proximity is correlated with network security risks. Most of these studies have largely focused on discovering the relevant social, economic and technological indicators of a country that influence global cybersecurity. However, none of these studies have leveraged this into the context of devising a predictive cybersecurity model that can potentially be useful in detecting and preventing cyberattacks.

2.3 Reputation Scoring

The concept of reputation refers to an opinion or belief held about something. A number of studies have proposed approaches for reputation scoring of IPs in computer networks. In order to deduce the reputation scores, features are compared against those of known signatures [22, 23, 24, 25]. For example, Renjan et al. [25] propose a vector space approach that uses euclidean distance to compare features of an IP address to those of blacklisted IPs. Interestingly, Antonakakis et al. proposes a dynamic approach to reputation scoring that adjusts the reputation score according to the level of maliciousness [24].

On the other hand, [26, 27], highlight limitations associated with dependence on blacklists, where attack signatures may change and thus render blacklists inefficient. [26] describes IP reputation as a concept of rating a host based on their past actions and comparing high level information such as domain names to a group of hosts whose reputation is known. Coskun [28] proposes a clustering approach to detect groups of hosts within a network based on whether traffic patterns are associated with malicious or benign activity. To the best of our knowledge, the existing studies have focused on network related features and neglect the role of geo-contextual features in assessing the reputation of a host based on its geo-related characteristics.

3 Methodology

Our overall approach is categorized in two phases as illustrated in Figure 2:

1) Discovery phase which describes data curation, pre-processing, and baseline model generation.
2) IP reputation scoring phase which describes extraction of network and geo-contextual information for an IP address and evaluation of the reputation scoring model.

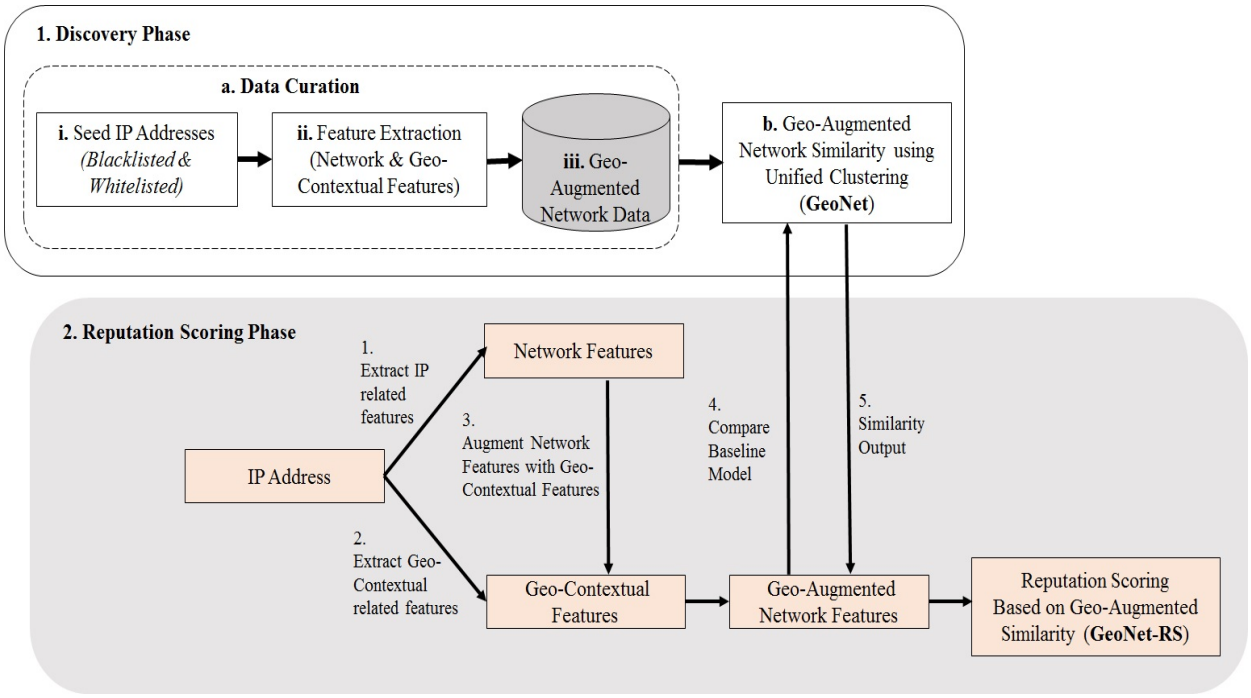

Figure 2: System Architecture

Specifically, 2.1 (a), data curation is applied whereby seed IP addresses are collected in 2.1(i). This is followed by the extraction of network and geo-contextual features matching the seed IPs in Figure 2.1(ii) and combined to generate a geo-augmented network dataset in 2.1(iii). A baseline model is then generated using unified clustering where geo-augmented network similarity is determined among seed IP addresses in 2.1 (b). This is referred to as the discovery phase. We anticipate that the discovery phase can be useful in offline mode for static IP analysis using historical data.

Then in Figure 2.2, an IP address is ingested for which network and geo-contextual features are extracted and augmented to form a geo-augmented network feature set. The combined feature set is evaluated against the baseline model to generate a similarity match. Finally, a reputation score is determined for the given IP address. This is referred to as the reputation scoring phase. We anticipate that the discovery phase can be useful in an online mode for real-time evaluation of incoming IP addresses. We discuss these steps in detail.

3.1 Data Curation

During data curation, seed IP data is collected from various open proprietary IP seed data sources. This data consists of labelled IP addresses which are either blacklisted or whitelisted. An IP address is said to be blacklisted if it has a history of exhibiting malicious activity, else it is considered whitelisted. Table 1 provides a summary of seed IP data sources. It should be noted that our data sources are credible specifically because they update the lists regularly. Therefore, the status of each IP address was valid by the time of this study. In the future, we plan to extend this study to detect whitelisted IPs that are linked to malicious behavior such as installing additional malware, among other anomalous behavior.

For each seed IP address, its network features are also extracted to create what we refer to as a network data set consisting of categorical features. The network data for each IP is mapped and integrated across the different data sources giving a combined total of 59 attributes for which 41 attributes are unique across all the network datasets. Table 1 provides a summary of network data sources (*see Appendix A.1 for full list of features*).

Following this, geo-contextual features are extracted from multiple city-level datasets from United Nations Statistics Division (UNSD) [35, 36, 37, 38, 39, 40, 41, 42] to generate a geo-contextual feature set consisting of predominantly continuous features. Our argument is that geo-contextual features are useful when there is limited session data such that a blacklisted IP address operates in a stealth mode. Therefore, we augment network data with geo-contextual features from multiple independent sources to increase and provide additional information to further characterize IP addresses. We compare the our proposed approach for geo-augmented network data to network data. Table 1 provides a summary of the geo-contextual data sources (*see Appendix A.2 for full list of features*).

Table 1: Summary of Data Sources

Data Source	Dataset Name	Open/Paid	Number of Features
A. IP Seed Data Sources	Blacklisted IPs: Cisco Talos [43, 44]	Open	NA (Seed data consists of the list of blacklisted and whitelisted IP addresses without IP-related features)
	Blacklisted IPs: CINS Score [45]	Open	
	Blacklisted IPs: MyIP [46]	Open	
	Blacklisted IPs: FireHol [47]	Open	
	Blacklisted IPs: Greensnow [48]	Open	
	Whitelisted IPs: OpenDNS [49]	Open	
B. Network Data Sources	Maxmind ASN & City Lite Databases [50]	Paid	10
	Maxmind GeoIP2 Precision - Insights [51]	Open	26
	Shodan [52]	Open	17
	Python Lib [53]	Open	5
	SpamHaus [54]	Open	1
	Total Network Features	-	**59 (*41 Unique)**
	Gender dataset [35]	Open	3
	Communication Infrastructure dataset [36]	Open	8

C. Geo-Contextual Data Sources	Water System dataset [37]	Open	5
	Fuel dataset [38]	Open	13
	Housing dataset [39]	Open	11
	Toilet dataset [40]	Open	17
	Waste Disposal dataset [41]	Open	12
	Living Quarters dataset [42]	Open	7
	Total Geo-Contextual Features		**76**

Using semantic mapping, the geo-contextual and network data are horizontally appended based on a common geographic feature, specifically 'city' that exists in both data representations. For example, given the city 'Delhi' in the geo-contextual dataset and 'New Delhi' in network dataset, the latter is semantically mapped to the former based on the city feature. The result is a network dataset augmented with geo-contextual features referred to as a geo-augmented network dataset. Hence, this consists of a combined feature vector of both continuous and categorical features. In this study, we evaluate our proposed model for network intrusion detection in the presence and absence of geo-contextual data.

3.2 Data Preprocessing

First, we identify missing values for which we apply statistical measures specifically, mean. For cases where a specific feature with missing values does not provide enough data to deduce missing values, we apply the values of associated features, particularly for geo-contextual features. For example, in order to determine the number of females in New York, we utilize the number of females in USA given that New York is geographically located within USA. On the other hand, we ignore any data instances where the city in the network dataset cannot be deduced or matched with the geo-contextual feature set.

We then apply feature transformation on each categorical feature (found in network feature set) using one-hot encoding where a categorical feature is transformed into a binary feature vector by expanding feature values. For example, let us consider the categorical feature - 'country' with possible values USA, India or China. The country feature is transformed into three binary features representing each country value such that an IP address will have a 'yes' or 'no' value for USA, India and China respectively.

Additionally, data transformation using a common naming convention is applied. For example, given a city name 'Delhi' versus 'New Delhi', we assign a common name 'Delhi'. Then, for all continuous features, we apply min-max normalization to generate data values within 0 to 1 range. We also apply equal frequency binning to ensure a class balance between blacklisted and whitelisted IPs.

Furthermore, the augmentation of network and geo-contextual data, and data transformation of categorical features into binary vectors, increases the feature space, thus generating high dimensional data. Thus, feature selection is applied specifically using pearson's correlation which is a bivariate analysis that measures the strength of association between any selected feature and the class label (blacklisted/whitelisted). It should be noted that for the categorical variables that are converted to binary attributes (using one-hot encoding), an overall correlation for binary attributes is arrived at via a weighted average [29]. Overall, attributes are ranked based on their individual evaluations from which a subset of top ranked attributes are selected. In Table 4 (section 4), we provide a summary of the count of geo-augmented features before and after feature selection.

We provide an illustration of our data pre-processing step in the following walk-through example.

Walk-Through Example:

First, we explain the data pre-processing journey that a typical raw network dataset undergoes before it is ready for analysis in Figure 3.

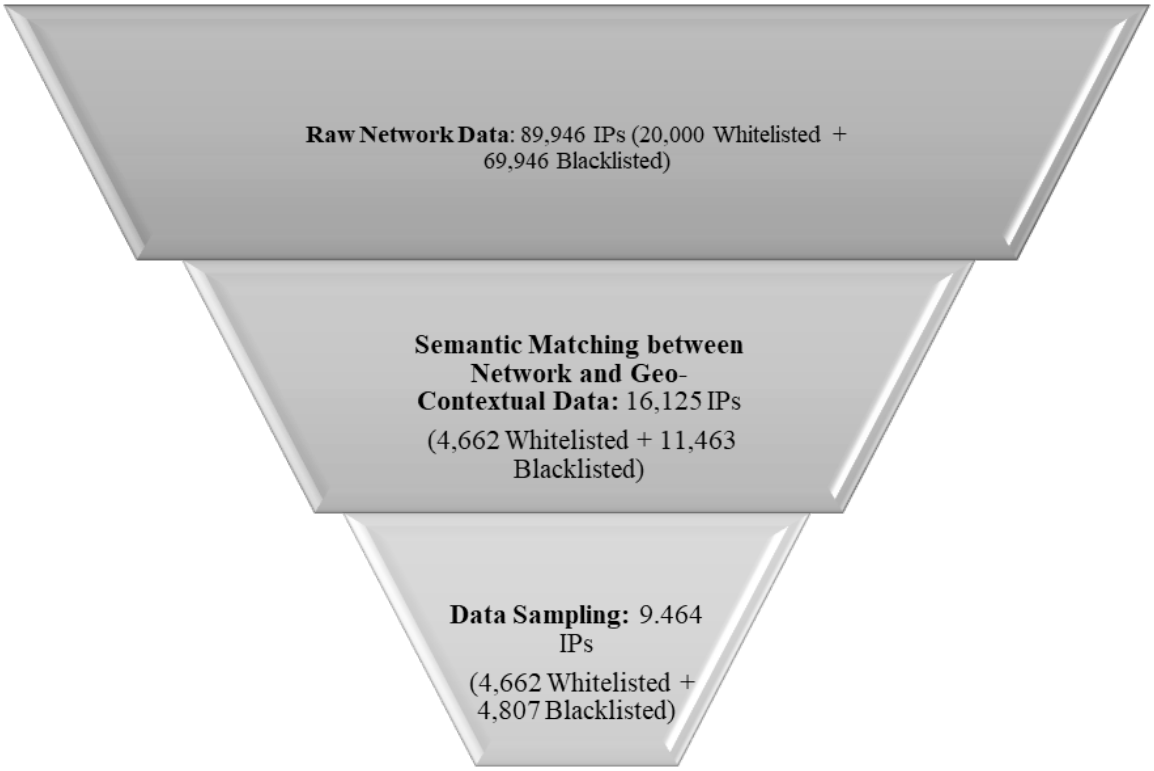

Figure 3: A Journey of Record Pre-Processing

Given a network dataset in Figure 3 which consists of the following;

- 89,946 seed IP Addresses - 20,000 whitelisted IPs and 69,946 blacklisted IPs.
- Each IPs is matched to a set of geo-contextual features based on the city. Out of 89,946 IPs, only 16,125 IPs have a city match between the network dataset and geo-contextual dataset. This therefore generates a geo-augmented network dataset with only 16,125 IPs out of 89,946 IPs that we first began with.
- Given 16,125 IPs - 4,662 whitelisted and 11,463 blacklisted IPs, thus indicating an imbalanced number of blacklisted and whitelisted IPs. In order to minimize class bias, random sampling is applied on the 16,125 IPs to obtain a balanced number of whitelisted and blacklisted IPs.
- The resulting sample consists of 9,469 IPs - 4,662 whitelisted and 4,807 blacklisted IPs to be utilized for analysis.

Next, we explain feature pre-processing for a geo-augmented network dataset consisting of the final selection of whitelisted and blacklisted IPs in Figure 4.

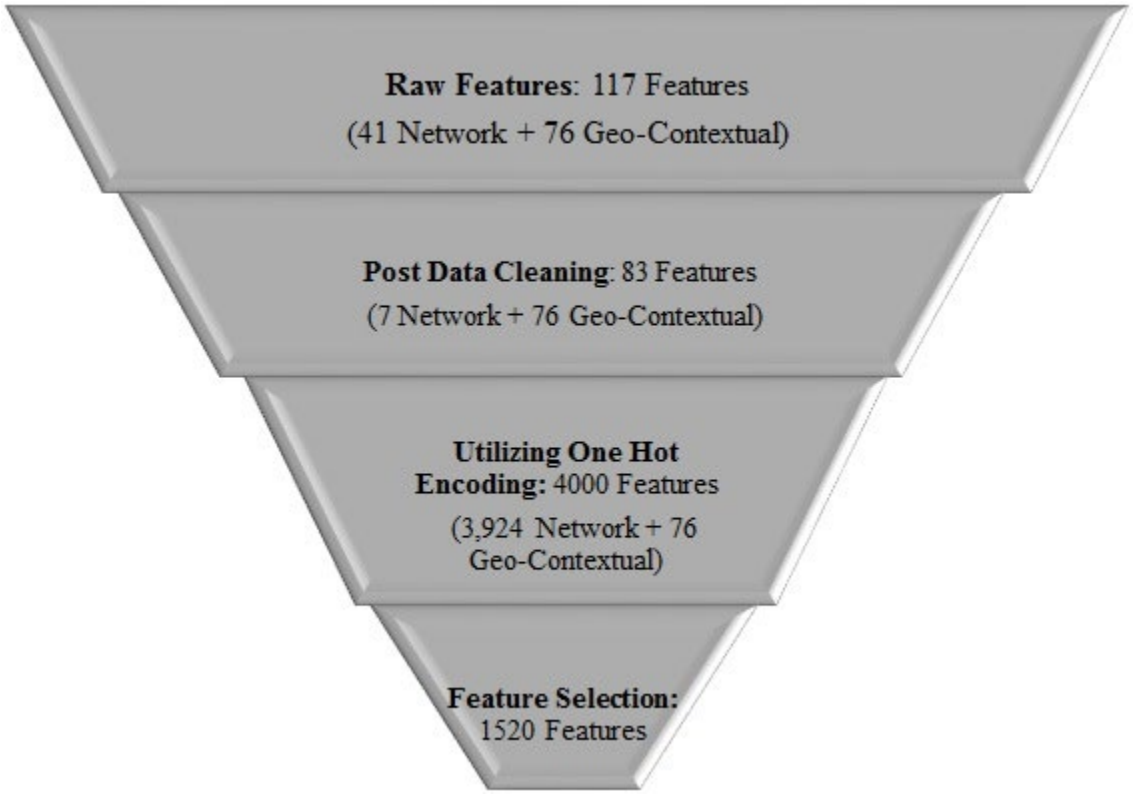

Figure 4: A Journey of Feature Pre-Processing

Given a geo-augmented network dataset in Figure 4 which consists of the following;

- There are 117 raw data features – 41 network features and 76 geo-contextual features.
- Following usability, sparsity and redundancy checks, 83 features are selected - 7 network features and 76 geo-contextual features.
- Given all the network features are categorical, the 7 network features are converted into binary features using one-hot encoding thus generating 3,924 network features.
- Then given all the geo-contextual features are continuous, normalization is performed to generate values within 0 to 1 range. The resulting feature set consists of 4,000 features - 3,924 network features and 76 geo-contextual features.
- Feature selection is applied on the 4,000 features using pearson's correlation, and thus generating a feature subset of approximately 1,520 features which is based on their ranked evaluation of each feature to the class feature (blacklisted/whitelisted).

3.3 Geo-Augmented Network Similarity (GeoNet) using Unified Clustering

This study proposes unified clustering where we apply clustering to heterogeneous features (both continuous and categorical) in the geo-augmented network dataset using a unified distance to determine the clustering cost using a dissimilarity score between data objects. Our proposed approach leverages [6, 31], where they determine a similarity score.

First, we determine the similarity *catSim* between categorical features using jaccard coefficient as:

$$\text{catSim}\,(i, j) = \frac{|\, i \cap j \,|}{|i \cup j|} \tag{1}$$

Next, we determine the numeric distance *numDist* between continuous features using euclidean distance as:

$$\text{numDist (i, j)} = \sqrt{\sum_{q}^{p} (i_q - j_q)^2} \tag{2}$$

We then convert the categorical similarity to a categorical distance *catDist* as:

$$\text{CatDist (i, j)} = (1 - \text{catSim}) \tag{3}$$

Finally, we aggregate the categorical and continuous data distances to generate the dissimilarity score *d(i, j)* as:

$$\text{d(i, j)} = \text{numDist(i, j)} + \text{catDist (i. j)} \tag{4}$$

Next, using the elbow method to determine Ideal *k*, where *k* is the number of clusters [32, 33]. The elbow method is a well-known decision approach based on the observation that increasing the number of clusters can help to reduce the sum of within cluster variance of each cluster [32]. In this study, we apply the clustering cost, which is the overall dissimilarity score for the clustering to determine *k*. In the future, we plan to extend this study to explore other methods such as silhouette coefficient, gap statistic and cross validation to determine the number of clusters. Algorithm 1 describes unified clustering in which the initial cluster centroids are randomly selected. We determine cluster assignments iteratively based on the dissimilarity score between IP addresses and the cluster centroids to which we assign each IP address to a cluster with the minimum dissimilarity score (see equation 4).

ALGORITHM 1: Unified Clustering

Input: ***n***: total number of instances.
k: number of clusters to be formed.
IPList: List of IPs $\{i_1, i_2, \ldots i_n\}$ containing continuous and categorical attributes.
centroidList: initial list of cluster centroids $\{c_1, c_2 \ldots c_k\}$
clusterList: List of clusters $\{C_1, C_2, \ldots C_k\}$

Require: $k >= 2$

For all instance i_i in instancesList, do
  For all centroid c_i in centroidList, do
    Determine dissimilarity score d between i_i and c_i.
    Assign i_i to the cluster such that d_i is minimum.
  End for
End for
Repeat
For all clusters C_i recalculate centroid, do
  Reassign the data instances to new clusters based on their original and new dissimilarity scores
End for

Overall, cluster evaluation is based on two intrinsic methods, that is; i) clustering cost, which measures the combined dissimilarity scores for all clusters and ii) silhouette coefficient, which measures the cohesiveness and separation of clusters [32]. However, in both evaluation metrics, it is important to note that the distance measure applied is based on the dissimilarity score (see equation 4) in order to account for both continuous and categorical attributes.

Our proposed model utilizes geo-augmented network data to detect similarity in IP behavior, a concept we refer to as **Geo-Augmented Network (GeoNet) Similarity using Unified Clustering.** Prior studies using unified

clustering demonstrate promising results for datasets with heterogeneous features [6, 30, 31]. We also compare unified clustering to k-prototype [15], a method commonly utilized in clustering heterogeneous feature sets containing categorical and continuous attributes. For comparison to network data, which consists only of categorical attributes, we compare against k-modes [11].

Furthermore, given that the data sources used in this study consist of labelled data, and therefore provide ground truth, we also validate the cluster outcomes using; i) cluster homogeneity, which measures the purity of a cluster in a clustering, ii) cluster completeness, which requires that a clustering should assign objects belonging to same category (in ground truth) to the same cluster, and, iii) v-measure, which measures the harmonic mean between cluster homogeneity and completeness [34].

3.4 IP Reputation Scoring based on Geo-Augmented Network Similarity (GeoNet-RS)

Based on clustering outcomes, we compare an IP address to determine the probability of it being blacklisted or otherwise. For this, the reputation for a given IP address is determined by generating a reputation score between 0-1 where 1 is the highest probability of being blacklisted and 0 otherwise. In this study, we conduct a reputation analysis based on five reputation scoring methods, where each method employs varying parameters to determine the reputation score. Our proposed approach for IP reputation scoring is referred to as **IP Reputation Scoring based on Geo-Augmented Network Similarity (GeoNet-RS).** We anticipate that the discovery phase can be useful in an online mode for real-time evaluation of incoming IP addresses.

Each reputation scoring method takes the following baseline inputs:

- ***k***: Number of clusters
- ***IPList*:** IP addresses $\{i_1, i_2, \ldots i_n\}$ for which the reputation score needs to be calculated.
- ***blackProbabilityList***: List of each cluster's black probability $\{b_1, b_2, \ldots b_k\}$ where black probability identifies the composition of blacklisted IPs in a given cluster. For example, if a given cluster c_i has 40% blacklisted IPs, and then its blackProbability b_i is 0.40.
- ***centroidList:*** List of each cluster's centroid $\{c_1, c_2, \ldots c_k\}$. The scoring model calculates the distance between a given IP instance i_1 from each of the cluster centroids $\{c_1, c_2, \ldots c_k\}$ to create a ***dissimilarityList.***
- ***dissimilarityList:*** List of dissimilarity scores $\{d_1, d_2, \ldots d_k\}$ of a given IP instance i_1 from each of the cluster centroids.

Consider a scenario shown in Figure 5 that further explains the inputs.

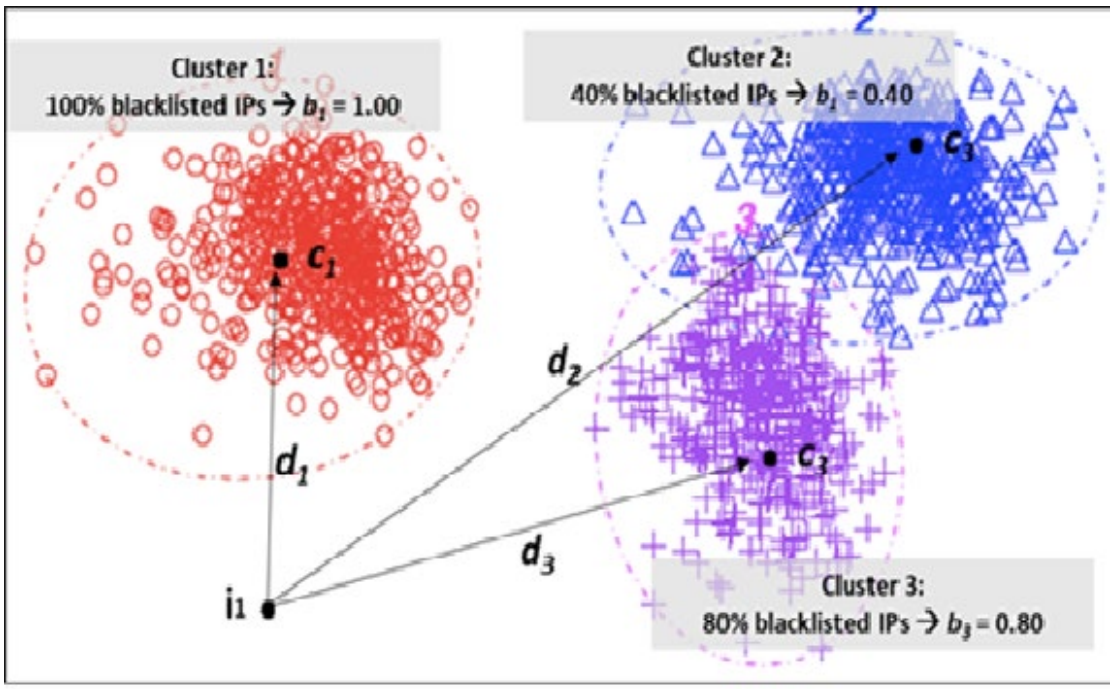

Figure 5: Reputation Scoring Scenario

Assume k is 3 where k is the total number of clusters generated from the clustering phase. i_1 is an IP instance for which the reputation score is to be determined. Given that cluster c_1 has 100% blacklisted IPs and thus black probability b_1 is 1. The centroid of cluster 1 *is* c_1. The scoring model will calculate the distance d_1 of i_1 from c_1. Similarly, given c_2 has 40% blacklisted IPs and thus b_2 is 0.40. The centroid of cluster 2 *is* c_2. The scoring model will calculate distance d_2 of i_1 from c_2. Finally, the respective values are determined for cluster 3. For this example, ***blackProbabilityList*** is $\{b_1, b_2, b_3\}$, ***centroidList*** is $\{c_1, c_2, c_3\}$ and ***dissimilarityList*** is $\{d_1, d_2, d_3\}$.

Using the baseline inputs above, we describe the proposed five reputation scoring methods along with their algorithmic representations. In Table 2, we provide an overview of the scoring methods as used in this study, followed by a description of each method.

Table 2: Overview of Reputation Scoring Methods

Algorithm	Method	Description
1	GeoNet-RS-C	Black Probability of Closest cluster
2	GeoNet-RS-AC	Average of Black Probabilities of ***x*** Closest clusters
3	GeoNet-RS-PC	Product of Black Probabilities of ***x*** Closest clusters
4	GeoNet-RS-WB	Weighted Average based on Blacklisted Probability
5	GeoNet-RS-WD	Weighted Average based on Distance

Black Probability of Closest Cluster

Algorithm 2 illustrates that given k clusters, the reputation score of an IP instance i_i is the black probability b_i of the closest cluster, where the closest cluster is one whose centroid c_i has the shortest distance d_i from i_i.

ALGORITHM 2: Black Probability of Closest Cluster (GeoNet-RS-C)

Input: ***k***: Number of clusters
***IPList*:** IP addresses $\{i_1, i_2, \ldots i_n\}$

***blackProbabilityList**:* List of each cluster's black probability $\{b_1, b_2, \ldots b_k\}$
centroidList: List of each cluster's centroid $\{c_1, c_2, \ldots c_k\}$.
dissimilarityList: List of dissimilarity scores $\{d_1, d_2, \ldots d_k\}$ of a given IP instance i_1 from each of the cluster centroids.

Output: Reputation Score s for each IP address i_i.

For each i_i in IPList, do
Extract network features
Add the geo-contextual features to the network features to form geo-augmented network feature vector i_i
Determine dissimilarity score d_i between i_i and each c_i in centroidList
Find minimum dissimilarity score min(d_i) to i_i.
Set Reputation score s = black probability b_i for c_i with min(d_i) to i_i
End for
Return Reputation Score s

Average of Black Probabilities of x Closest Clusters

Algorithm 3 illustrates that given k clusters, the reputation score of an IP instance i_i is the average of the black probabilities $\{b_1, b_2, \ldots b_x\}$ of x closest clusters where x closest clusters are the clusters whose centroids c_i have the shortest distance d_i from i_i. x is determined empirically by iterating over different values of x ranging from 1 to n and is thus user defined.

ALGORITHM 3: Average of Black Probabilities of x Closest Clusters (GeoNet-RS-AC)

Input: ***k***: Number of clusters
***IPList*:** IP addresses $\{i_1, i_2, \ldots i_n\}$
***blackProbabilityList**:* List of each cluster's black probability $\{b_1, b_2, \ldots b_k\}$
centroidList: List of each cluster's centroid $\{c_1, c_2, \ldots c_k\}$.
dissimilarityList: List of dissimilarity scores $\{d_1, d_2, \ldots d_k\}$ of a given IP instance i_1 from each of the cluster centroids.

Output: Reputation Score s for each IP address i_i.

For each i_i in IPList, do
Extract network features
Add the geo-contextual features to the network features to form geo-augmented network feature vector i_i
Determine dissimilarity score d_i between i_i and each c_i in centroidList
Given x closest clusters, where x ranges from *1 to n*
For each x,
Get average of black probabilities of x closest clusters. Iterate over different values of x.
Set Reputation score s = average black probability b_x for c_x with min(d_x) to i_i
End for
End for
Return Reputation Score s

Product of Black Probabilities of x Closest Clusters

Algorithm 4 illustrates that given k clusters, the reputation score of an IP instance i_i is the product of the black probabilities $\{b_1, b_2, \ldots b_x\}$ of x closest clusters. This approach is essentially similar to that described in Algorithm 3, but unlike the average, it takes on product of the black probabilities of the x closest clusters.

ALGORITHM 4: Product of Black Probabilities of x Closest Clusters (GeoNet-RS-PC)

Input: ***k***: Number of clusters
***IPList*:** IP addresses $\{i_1, i_2, \ldots i_n\}$
blackProbabilityList*:* List of each cluster's black probability $\{b_1, b_2, \ldots b_k\}$
centroidList: List of each cluster's centroid $\{c_1, c_2, \ldots c_k\}$.
dissimilarityList: List of dissimilarity scores $\{d_1, d_2, \ldots d_k\}$ of a given IP instance i_1 from each of the cluster centroids.

Output: Reputation Score s for each IP address i_i.

For each i_i in IPList, do

Extract network features

Add the geo-contextual features to the network features to form geo-augmented network feature vector i_i

Determine dissimilarity score d_i, between i_i and each c_i in centroidList

Given x closest clusters, where x ranges from *1 to n*

For each x,

Calculate product of black probabilities of x closest clusters

Set Reputation Score s = product of black probability b_x for c_x with min(d_x) to i_i

End for

End for

Return Reputation Score s

Weighted Average based on Blacklisted Probability

Algorithm 5 illustrates that given k clusters, the weighted average based on blacklisted probability $w(b)$ of an IP instance i_i from each cluster's centroid is determined given by:

$$w(b) = \frac{\sum_1^k (b * d)}{\sum_1^k b} \quad (5)$$

For example, given a dissimilarity score d_i of i_i from c_i is 0.53 and b_i for the same cluster is 0.75. Thus, the weighted probability for c_i is defined as $b_i * d_i$ which in this case is 0.53 * 0.75 = 0.40. We take products when things are mutually independent. In this case we argue that the probability of an IP being blacklisted and its distance from each centroid is assumed to be independent. In a more complex system dependencies would be mapped but that is not in the scope of this work and for simplicity we assume independence. Similarly, $b_i * d_i$ is calculated for all other clusters and the products are summed up to generate a weighted sum to determine $w(b)$.

ALGORITHM 5: Weighted Average based on Black Probability (GeoNet-RS-WB)

Input: ***k***: Number of clusters
***IPList*:** IP addresses $\{i_1, i_2, \ldots i_n\}$
blackProbabilityList*:* List of each cluster's black probability $\{b_1, b_2, \ldots b_k\}$

centroidList: List of each cluster's centroid $\{c_1, c_2, \ldots c_k\}$.
dissimilarityList: List of dissimilarity scores $\{d_1, d_2, \ldots d_k\}$ of a given IP instance i_1 from each of the cluster centroids.

Output: Reputation Score s for each IP address i_i.

For each i_i in IPList, do
    Extract network features
    Add the geo-contextual features to the network features to form geo-augmented network feature vector i_i
    Determine dissimilarity score d_i between i_i and each c_i in centroidList
    For each cluster
        Calculate Weighted Average based on Black Probability $w(b)$
        Set Reputation Score s = Weighted Average $w(b)$
    End for
End for
Return Reputation Score s

Weighted Average based on Distance
Algorithm 6 illustrates that given k clusters, the weighted average based on distance $w(d)$ from each cluster's centroid is determined as:

$$\mathrm{w(d)} = \frac{\sum_1^{\mathrm{k}}(\mathrm{b} * \mathrm{d})}{\sum_1^{\mathrm{k}} \mathrm{d}} \tag{6}$$

This method is similar to that described in Algorithm 5, however the denominator is based on the distances from cluster centroids.

ALGORITHM 6: Weighted Average based on Distance (GeoNet-RS-WD)

Input: ***k***: Number of clusters
IPList**:** IP addresses $\{i_1, i_2, \ldots i_n\}$
blackProbabilityList*:* List of each cluster's black probability $\{b_1, b_2, \ldots b_k\}$
centroidList: List of each cluster's centroid $\{c_1, c_2, \ldots c_k\}$.
dissimilarityList: List of dissimilarity scores $\{d_1, d_2, \ldots d_k\}$ of a given IP instance i_1 from each of the cluster centroids.

Output: Reputation Score s for each IP address i_i.

For each i_i in IPList, do
    Extract network features
    Add the geo-contextual features to the network features to form geo-augmented network feature vector i_i
    Determine dissimilarity score d_i between i_i and each c_i in centroidList
    For each cluster
        Calculate Weighted Average based on dissimilarity Score $w(d)$
        Set Reputation Score s = Weighted Average $w(d)$
    End for
End for

Return Reputation Score *s*

In this study, we evaluate each of the proposed reputation scoring methods based on ground truth utilizing accuracy, precision and recall as defined in equations 7, 8 and 9 respectively:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

Hence, we compare the predicted label to the actual label. Next, we present our findings.

4 Experimental Results

For our analysis of findings based on the proposed methods is as follows.

4.1 Datasets

In Table 3, we provide an overview of five network datasets generated for experimental purposes from the various seed IP and network data sources in Table 1. For purposes of our study, we create different sample datasets as shown in Table 3. It should be noted that some sample datasets are generated by combining the data sources 1, 2, 3, & 4 as indicated in Table 3. Figure 3 provides a detailed description on the data pre-processing. Given that our study proposes network intrusion in a geo-contextual context, we generate 4 geo-augmented network datasets. We also evaluate our proposed approach on imbalanced and balanced datasets.

Table 3: Summary of Network Datasets

Dataset Name	Source	Original Number of IPs	Selected Number of IPs	% of Blacklisted IPs	% of Whitelisted IPs
BW-Sampled, Balanced	(1) Maxmind GeoIP2 Precision - Insights (web service); (2) Shodan; (3) SpamHaus; (4) Python	8,136	370	51%	49%
BW-Sampled, Imbalanced		8,136	706	94%	6%
BW-Sampled, Large, Open		90,000	5,428	51%	49%
BW-Sampled, Large, Paid	Maxmind ASN & City Lite Databases	90,000	9,469	51%	49%

Table 5 provides a summary of the count of features with geo-augmentation before and after feature selection is applied with respect to each dataset. The walk-through example in Figure 4 provides detailed description on the feature pre-processing.

Table 5: Summary of Feature Counts for Geo-Augmented datasets

	Total Geo-Augmented Network Feature Set	Selected Features (based on Pearson's Correlation)
BW-Sampled, Balanced	661	159
BW-Sampled, Imbalanced	2150	304
BW-Sampled, Large, Open	2644	465
BW-Sampled, Large, Paid	4029	1503

For evaluation, we compare GeoNet using unified clustering and traditional clustering techniques, specifically k-modes and k-prototype as discussed in our methodology. Table 6 provides a summary of the clustering models with respect to each dataset.

Table 6: Summary of Clustering Models Evaluated in the Study

Dataset Name	Technique	Network Data	Geo-Augmented Network Data	Number of Models
BW-Sampled, Balanced	K-Modes	x		**3**
	K-Prototype		x	
	GeoNet		x	
BW-Sampled, Imbalanced	K-Modes	x		**3**
	K-Prototype		x	
	GeoNet		x	
BW-Sampled, Large, Open	K-Modes	x		**3**
	K-Prototype		x	
	GeoNet		x	
BW-Sampled, Large, Paid	K-Modes	x		**3**
	K-Prototype		x	
	GeoNet		x	
Overall Total		**4**	**8**	**12**

4.2 Determining Ideal *k*

Figure 6 illustrates selection of optimal number of clusters (*k*) obtained using the elbow method as discussed in the methodology section.. We ran several experiments using the elbow method and found consistent results.

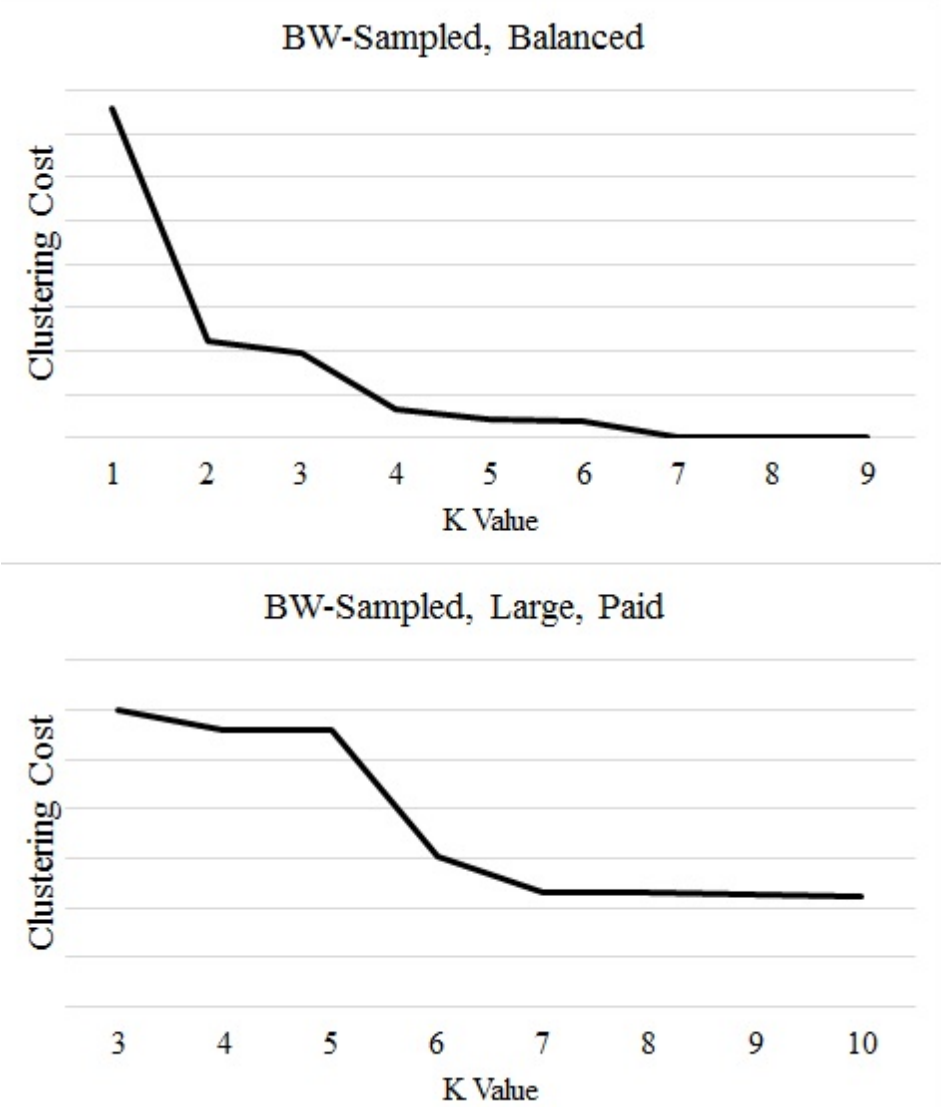

Figure 6: Determining k using Elbow Method

In Table 7, we provide a summary of *k* values per dataset.

Table 7: Summary of Optimal *k* Values Per Dataset

Dataset	***k* value**
BW-Sampled, Balanced	4
BW-Sampled, Imbalanced	4
BW-Sampled, Large, Open	7
BW-Sampled, Large, Paid	7

4.3 Comparison of Geo-Augmented Network Data versus Network Data

For our evaluation, we compare cluster outcomes for network datasets (without geo-contextual features) which consist of categorical features only and geo-augmented network datasets (with geo-contextual features) which consist of both categorical and continuous features. For the former we apply k-modes and for the latter we apply the proposed GeoNet similarity using unified clustering. Our evaluation is based on clustering cost and silhouette coefficient. Table 8 provides a summary of our findings.

Table 8 Evaluation of GeoNet (*Geo-Augmented Network Data)* versus K-Modes *(Network Data)* **Best Result is denoted with Bold

	Clustering Cost	Silhouette Coefficient
BW-Sampled-Balanced		
K-Modes	323	0.78
GeoNet	**18.51**	**0.86**
BW-Sampled-Imbalanced		
K-Modes	1489	0.67
GeoNet	**43.87**	**0.85**
BW-Sampled Large Open		
K-Modes	11369	0.30
GeoNet	**61.56**	**0.95**
BW-Sampled Large Paid		
K-Modes	14574	0.21
GeoNet	**310**	**0.93**

Our findings in Table 8 shows that augmenting network data with geo-contextual features results into significantly better clusters as indicated by the lower clustering cost compared to cases without geo-augmentation. This is observed across all datasets where GeoNet outperforms k-modes. Similarly, we observe cohesive and well separated clusters with GeoNet as indicated by a higher silhouette coefficient. This indicates that augmenting additional contextual information increases the likelihood for identifying similarity among IP instances. Overall, our findings clearly demonstrate that examining geo-contextual perspectives improves cluster outcomes required for effective intrusion detection. Hence, we claim the following:

Claim I. *Augmenting the network datasets with geo-contextual features improves the clustering outcomes*

4.4 Comparison of GeoNet versus K-Prototype for Geo-Augmented Network Data

We compare our findings for GeoNet with k-prototype to evaluate specifically their performance on clustering heterogeneous features based on clustering cost and silhouette coefficient respectively. Table 9 provides a summary of our findings.

Table 9 Evaluation of GeoNet versus K-Prototype for Geo-Augmented Network Data
***Best Result is denoted with* **Bold**

	Clustering Cost	Silhouette Coefficient
BW-Sampled-Balanced		
K-Prototype	56.71	0.65
GeoNet	**18.51**	**0.86**
BW-Sampled-Imbalanced		
K-Prototype	244.05	0.57
GeoNet	**43.87**	**0.85**
BW-Sampled Large Open		
K-Prototype	1102.85	0.46
GeoNet	**61.56**	**0.95**
BW-Sampled Large Paid		
K-Prototype	2774.84	0.32
GeoNet	**310**	**0.93**

Our findings in Table 9 shows that GeoNet using unified clustering significantly outperforms k-prototype clustering when evaluated on geo-augmented network data across all datasets. This is portrayed by the significantly lower clustering costs generated thus forming more compact clusters. In the same way, we observe cohesive and well-separated clusters with GeoNet as indicated by a higher silhouette coefficient. This further indicates that GeoNet using unified clustering handles large heterogeneous data well. This is therefore useful in detecting and assessing cyberthreats in massive datasets. For this, we make the following claim:

Claim II. *For datasets with heterogeneous features, Unified Clustering is the best baseline model.*

4.5 Ground Truth Evaluation for Geo-Augmented Network based Similarity

We also evaluate cluster outcomes for all models using cluster homogeneity, completeness and v-measure to evaluate the purity and inclusiveness of clusters [34]. Our evaluation is based on ground truth, that is the labeled instances of blacklisted and whitelisted IPs with respect to each dataset as it is utilized in this study.

Table 10: Comparative Analysis of Cluster Homogeneity, Completeness and V-Measure for Geo-Augmented Network Data versus Network Data
***Best Result is denoted with* **Bold**

	Homogeneity	Completeness	V-Measure
BW-Sampled-Balanced			
K-Modes	0.03	**0.074**	0.04
K-Prototype	**0.1**	0.066	**0.09**
GeoNet	0.02	0.06	0.03
BW-Sampled-Imbalanced			
K-Modes	**0.6**	**0.4**	**0.5**
K-Prototype	0.03	0.03	0.03

GeoNet	0.1	0.05	0.1
BW-Sampled Large Open			
K-Modes	**0.9**	**0.4**	**0.5**
K-Prototype	0.23	0.10	0.13
GeoNet	0.04	0.02	0.03
BW-Sampled Large Paid			
K-Modes	**0.6**	**0.2**	**0.4**
K-Prototype	0.06	0.07	0.063
GeoNet	0.09	0.04	0.058

Table 10 shows that overall k-modes which uses network data only, generates more homogeneous clusters when compared again the classes in the labelled datasets as indicated by higher homogeneity scores in the range of 0.6 to 0.9. In contrast, the clusters generated from augmenting geo-contextual data are less homogeneous clusters as indicated by lower homogeneity scores of <= 0.2 observed in both GeoNet and k-prototype. It is evident that the clusters formed by analyzing only network data are limited to traditional intrusion detection techniques, which are useful when one trusts the network session data. On the otherhand, the clusters formed by analyzing geo-augmented network data reveal concealed information about an IP address and thus pose potential in detecting underground attacks that operate in stealth mode such as advanced persistent threats (APTs).

These observations present interesting findings and thus demonstrate the applicability of our proposed approach in intrusion detection to serve as a supplement when there is not enough network session data to detect similarity in attack behavior. It can therefore be argued that our approach accounts for the context presented by geo-contextual features to detect cyber threats.

Further, other measures like completeness and v-measure are not an ideal evaluation criterion for our approach and others alike, particularly when one seeks to identify unusual behavior in the network that is captured in certain cohesive groups of IPs based on similarity in their behavior. This also results into well-separated clusters, which do not necessarily match up to the class labels. It is therefore a clear indicator that some instances may not fit well into a given class and thus portray anomalous behavior [55].

4.6 Ground Truth Evaluation for IP Reputation Scoring

In this section, we extend our analysis to evaluate the IP reputation scoring for clustered IPs particularly based on GeoNet. We evaluate all the 5 scoring methods proposed in our methodology section (see algorithms 2 to 6). We compare the prediction outcomes determined by the reputation scores to ground truth (using the class labels with respect to blacklisted IPs). For this, we measure the accuracy. Assume one of the test IP instances in the seed data is 153.134.130.54. This IP address is a blacklisted IP address. Therefore, the reputation score for this IP address will be considered accurate if it is able to predict that 153.134.130.54 is a blacklisted IP. To minimize redundancy, we present findings for only a single dataset, that is, the BW-Sampled, Large, Paid (*where k = 7*). Also, for the reputation scoring methods based on *x* closest clusters (see Algorithm 3 and 4 respectively), *x* is defined as 3 in this case. This is because with k = 7, *x* was tested incrementally from 1 to 7 and the 3 closest clusters provided the best prediction outcomes in terms of number of correct predictions as determined by the accuracy. Figure 7 illustrates our findings.

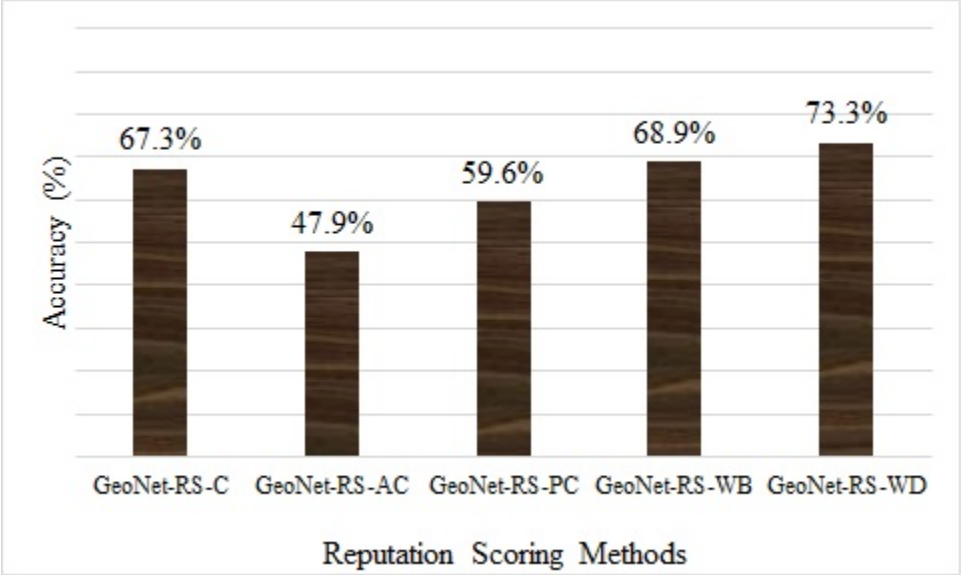

Figure 7: Evaluation of Reputation Scoring Methods (*using BW-Sampled, Large, Paid*)

Our findings in Figure 7 indicate that when examining IP instances, considering the weighted average based on the distance (**GeoNet-RS-WD**) produces the IP reputation score with the highest prediction outcome in terms of accuracy as described in our methodology. Our findings also indicate that weighted average based on the black probability (**GeoNet-RS-WB**) as well as examining of the closest cluster (**GeoNet-RS-C**) are also suitable in determining a relatively accurate IP reputation.
On the other hand, taking only a selected number of clusters deemed as closest results into a lower IP reputation score. This can be characterized by biased behavior of other IPs, which may not necessarily be similar.

We further analyze **GeoNet-RS-WD** to compare prediction outcomes for geo-augmented network data and network data. For this, we measure accuracy, recall and precision as illustrated in Figure 8.

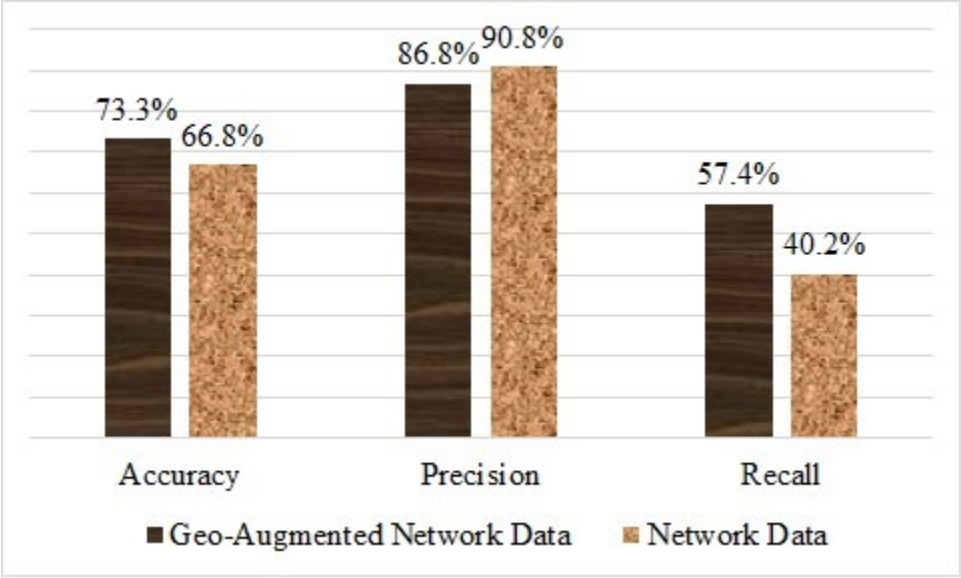

Figure 8: Evaluating Reputation Scoring on Geo-Augmented Network Data versus Network Data (*using BW-Sampled, Large, Paid*)

Our findings in Figure 8 clearly indicate that by augmenting network data with geo-contextual features the prediction of blacklisted and whitelisted IP addresses outperforms that with relies on only network features. This is specifically evident in the accuracy and recall. This is also further confirms our findings in Table 8, which illustrates that the clusters formed are more cohesive and well-separated.

5 Discussion

In this section, we present an overview of our results

First, the augmentation of network data with geo-contextual features detects higher similarity in the behavior of IP Addresses. This is clearly demonstrated by the formation of better clusters characterized by lower clustering cost and high silhouette coefficient compared to modelling IP behavior by utilizing network features only as summarized in Table 11. Therefore, the inclusion of geo-contextual features increases the perspective on characterizing threats and potentially arms decision makers with the potential to be proactive in detecting or predicting cyber-attacks.

Table 11: Summary of Comparison Contexts based on Cluster Performance

Cluster Performance		
	Clustering Cost	**Silhouette Coefficient**
Approach		
Network Data	⇩	⇩
Geo-Augmented Network Data	⇧	⇧
⇧ *Performed Highest based on Evaluation Metric*		
⇩ *Performed Lowest based on Evaluation Metric*		

Second, our proposed approach of GeoNet similarity using unified clustering outperforms clustering techniques like k-prototype, commonly utilized in heterogeneous datasets. This is clearly observed in the high intra-cluster similarity and low inter-cluster similarity. Therefore, GeoNet is well-suited to handle large heterogeneous files compared to k-prototype.

Third, the augmentation of network data with geo-contextual features identifies unusual groupings of IP addresses as depicted by lower homogeneity, completeness and v-measure scores in comparison to clusters formed in network data as summarized in Table 12. The proposed approach is particularly useful when network encryption is utilized, in order to detect atypical activity in cases where there is limited network session data. Unlike signature-based techniques, our approach poses merit in the detection of APTs and subversion attacks where compromised IPs fly under the radar for an extended period.

Table 12: Summary of Comparison Contexts based on Ground Truth

Ground Truth Evaluation			
	Homogeneity	**Completeness**	**V-Measure**
Approach			
Network Data	⇧	⇧	⇧
Geo-Augmented Network Data	⇩	⇩	⇩
⇧ *Performed Highest based on Evaluation Metric*			
⇩ *Performed Lowest based on Evaluation Metric*			

Finally, in order to determine a reputation score for IP addresses in terms of both network and geo-contextual knowledge, our proposed approach essentially allows for scoring IP addresses by evaluating IP similarity based on weighted average based on the distance (GeoNet-RS-WD), weighted average based on the black probability (GeoNet-RS-WB) as well as examining of the closest cluster (GeoNet-RS-C). By evaluating against ground truth, our best scoring algorithm (GeoNet-RS-WD) gives a prediction outcome in terms of accuracy of 73.3% and precision of 86.6%. Table 12 summarizes our findings. It should be considered that our proposed approach is applied to a massive feature set. Based on this, there is also a tradeoff to ensure accommodating a richer information set when network data is encrypted, where no additional information is available beyond the data header. In such scenarios, it becomes imperative to have some insights, which may not be otherwise available. Our approach fills that gap.

Table 13: Summary of Comparison Contexts based on Reputation Scoring

Reputation Scoring			
	Accuracy	**Precision**	**Recall**
Approach			
Network Data	⇩	⇧	⇩
Geo-Augmented Network Data	⇧	⇩	⇧
⇧ *Performed Highest based on Evaluation Metric*			
⇩ *Performed Lowest based on Evaluation Metric*			

6 Conclusion and Future Work

In this study, we propose a novel approach to assess the reputation of an IP address using geo-contextual features. Geo-contextual data can be linked to location-specific non-spatial data, which encompasses a region's economic, social, demographic, and technological domains. This study's approach utilizes unified clustering, a technique that overcomes the problems of using heterogeneous features, that is, both continuous and categorical features in clustering. We present extensive experimental results that highlight the importance of geo-contextual knowledge in explaining network anomalies and compare several traditional clustering methodologies with unified clustering approach. Thus, this study's contributions are three-fold. First, we show that the approach of combining traditional network features with geo-contextual features presents a more robust and unique representation of hosts on a

network; Second, this study provides an empirical validation of applying unified clustering with geo-augmented network data in the cybersecurity domain to characterize IP behavior. Third, we have devised a novel reputation scoring model for an IP address based on geo-augmented network similarity. Findings from this study have implications in anomaly detection for cybersecurity applications, especially when there is limited information about the network session or there is a lack of historical data for the network features.

In the future, we would like to expand the network feature set to include domain related features along with IP related features to make the network dataset more robust. We would also continue to improve the accuracy of the clustering model which in turn would improve the accuracy of the scoring algorithms by using iterative clustering, a technique that extends unified clustering to further breakdown any malformed clusters.
Additionally, for this study, we utilized a static geo-contextual dataset. However, for future studies, a more dynamic geo-contextual dataset is recommended that accounts for socio-economic changes across the global cities over time. It would be interesting to observe the changes in cyber threats with the changing socio-political environment.

REFERENCES

[1] Waldo R. Tobler. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography* 46, no. sup1, 234-240.

[2] Crowdstrike. 2018. 2018 CrowdStrike Global Threat Report: Blurring the Lines Between Statecraft and Tradecraft. *Retrieved March 18, 2018, from https://www.crowdstrike.com/resources/reports/2018-crowdstrike-global-threat-report-blurring-the-lines-between-statecraft-and-tradecraft/*

[3] David Burt, Paul Nicholas, Kevin Sullivan, and Travis Scoles. 2014. The Cybersecurity Risk Paradox.

[4] ESRI. 2015. The Geo-contextual Approach to Cybersecurity. *Retrieved September 7, 2017, from http://www.esri.com/library/whitepapers/pdfs/geo-contextual-approach-to-cybersecurity.pdf*

[5] Ghita Mezzour. 2015. Assessing the global cyber and biological threat.

[6] Anuja Kench, Vandana P. Janeja, Yelena Yesha, Naphtali Rishe, Michael A. Grasso, and Amanda Niskar. 2015. Clinico-genomic data analytics for precision diagnosis and disease management. In *2015 International Conference on Healthcare Informatics*, pp. 263-271. IEEE.

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, 96, 34, pp. 226-231.

[8] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. BIRCH: an efficient data clustering method for very large databases. *ACM Sigmod Record* 25, 2, 103-114.

[9] Ming-Yi Shih, Jar-Wen Jheng, and Lien-Fu Lai. 2010. A two-step method for clustering mixed categroical and numeric data. *Tamkang Journal of science and Engineering* 13, 1, 11-19.

[10] Henri Ralambondrainy. 1995. A conceptual version of the K-means algorithm. *Pattern Recognition Letters* 16, 11, 1147-1157.

[11] Zhexue Huang and Michael K. Ng. 1999. A fuzzy k-modes algorithm for clustering categorical data. *IEEE transactions on Fuzzy Systems* 7, 4, 446-452.

12] M.V. Jagannatha Reddy and Balli Kavitha. 2012. Clustering the mixed numerical and categorical dataset using similarity weight and filter method. *International Journal of Database Theory and Application* 5.1, 121-134.

[13] Zengyou He, Xiaofei Xu, and Shengchun Deng. 2005. Clustering mixed numeric and categorical data: A cluster ensemble approach. *arXiv preprint cs/0509011*.

[14] Damien McParland, Catherine M. Phillips, Lorraine Brennan, Helen M. Roche, and Isobel Claire Gormley. 2017. Clustering high-dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data. *Statistics in medicine* 36, 28, 4548-4569.

[15] Zhexue Huang. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery* 2, 3, 283-304.

[16] Ishak Boushaki Saida, Kamel Nadjet, and Bendjeghaba Omar. 2014. A new algorithm for data clustering based on cuckoo search optimization. *Genetic and Evolutionary Computing*. Springer, Cham, 55-64.

[17] R. Madhuri, M. Ramakrishna Murty, J. V. R. Murthy, PVGD Prasad Reddy, and Suresh C. Satapathy. 2014. Cluster analysis on different data sets using K-modes and K-prototype algorithms. In *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol II*, pp. 137-144. Springer, Cham.

[18] Khadija Musayeva, Tristan Henderson, John BO Mitchell, and Lazaros Mavridis. PFClust: an optimised implementation of a parameter-free clustering algorithm. *Source code for biology and medicine* 9, 1, 5.
[19] Hong Jia, and Yiu-Ming Cheung. 2017. Subspace clustering of categorical and numerical data with an unknown number of clusters. *IEEE transactions on neural networks and learning systems* 29, 8, 3308-3325.
[20] Qiu-Hong Wang, and Seung Hyun KIM. 2009. Cyber attacks: Cross-country interdependence and enforcement. WEIS.
[21] Kumar and Carley, 2016. Sumeet Kumar, and Kathleen M. Carley. 2016. Approaches to understanding the motivations behind cyber attacks. In 2016 IEEE Conference on Intelligence and Security Informatics (ISI), pp. 307-309. IEEE.
[22] Yoshiro Fukushima, Yoshiaki Hori, and Kouichi Sakurai. 2011. Proactive blacklisting for malicious web sites by reputation evaluation based on domain and ip address registration. In 2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications, pp. 352-361. IEEE.
[23] Ashley Thomas. 2010. Rapid: Reputation based approach for improving intrusion detection effectiveness. In 2010 Sixth International Conference on Information Assurance and Security, pp. 118-124. IEEE.
[24] Manos Antonakakis, Roberto Perdisci, David Dagon, Wenke Lee, and Nick Feamster. 2010. Building a dynamic reputation system for dns. In USENIX security symposium, pp. 273-290.
[25] Arya Renjan, Karuna Pande Joshi, Sandeep Nair Narayanan, and Anupam Joshi. 2018. Dabr: Dynamic attribute-based reputation scoring for malicious ip address detection. In 2018 IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 64-69. IEEE.
[26] Wesley Pronk. 2011. Real-time Blacklisting of Bots based on Spam Analyis. In 15th Twente Student Conference on IT June 20th.
[27] Vaclav Bartos, Martin Zadnik, Sheikh Mahbub Habib, and Emmanouil Vasilomanolakis. 2019. Network entity characterization and attack prediction. Future Generation Computer Systems 97, 674-686.
[28] Baris Coskun. 2017. (Un) wisdom of Crowds: Accurately Spotting Malicious IP Clusters Using Not-So-Accurate IP Blacklists. IEEE Transactions on Information Forensics and Security 12, 6, 1406-1417.
[29] Jason Brownlee. 2016. How to perform feature selection with machine learning data in weka. *Retrieved April, 2020 from, https://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/*
[30] Vasundhara Misal, Vandana P. Janeja, Sai C. Pallaprolu, Yelena Yesha, and Raghu Chintalapati. 2016. Iterative unified clustering in big data. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3412-3421. IEEE.
[31] Vandana P. Janeja, Josephine M. Namayanja, Yelena Yesha, Anuja Kench, Vasundhara Misal. *International Journal of Data Warehousing and Mining (In press)*
[32] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. Data mining: concepts and techniques. Elsevier.
[33] Namayanja, Josephine, and Vandana P. Janeja. 2011. Subspace discovery for disease management: A case study in metabolic syndrome. *International Journal of Computational Models and Algorithms in Medicine (IJCMAM)* 2, 1, 38-59
[34] Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp. 410-420.
[35] United Nations Statistics Division. City population by sex, city and city type. Retrieved September 7, 2017, from http://data.un.org/Data.aspx?d=POP&f=tableCode:240#POP
[36] United Nations Statistics Division. Households in housing units by type of housing unit and availability of communication technology devices/access to Internet for selected cities. Retrieved September 7, 2017, from http://data.un.org/Data.aspx?q=city&d=POP&f=tableCode:307
[37] United Nations Statistics Division. Occupied housing units by type of housing unit and type of water supply system for selected cities. Retrieved September 7, 2017, from http://data.un.org/Data.aspx?q=city&d=POP&f=tableCode:283
[38] United Nations Statistics Division. Occupied housing units by type of housing unit and main type of fuel used for cooking for selected cities. Retrieved September 7, 2017, from http://data.un.org/Data.aspx?q=city&d=POP&f=tableCode:293
[39] United Nations Statistics Division. Occupied housing units by type of housing unit for selected cities. Retrieved September 7, 2017, from http://data.un.org/Data.aspx?q=city&d=POP&f=tableCode:279
[40] United Nations Statistics Division. Occupied housing units by type of housing unit and type of toilet for selected cities. Retrieved September 7, 2017, from http://data.un.org/Data.aspx?q=city&d=POP&f=tableCode:287

[41] United Nations Statistics Division. Occupied housing units by type of housing unit and main type of solid waste disposal for selected cities. Retrieved September 7, 2017, from http://data.un.org/Data.aspx?q=city&d=POP&f=tableCode:297
[42] United Nations Statistics Division. Living quarters by broad types for selected cities. Retrieved September 7, 2017, from http://data.un.org/Data.aspx?q=city&d=POP&f=tableCode:277
[43] Cisco Talos. Comprehensive Threat Intelligence. Retrieved January 5, 2018, from https://www.cisco.com/c/en/us/products/security/talos.html
[44] Cisco Talos. IP and Domain Reputation Center. Retrieved January 5, 2018, from www.talosintelligence.com/reputation
[45] Cins Score. Certtools/intelmq-feeds-documentation. Retrieved January 6, 2018, from https://github.com/certtools/intelmq-feeds-documentation/blob/master/CINSscore/Blacklist.md
[46] MyIP. (Real-Time). Blacklist IP Addresses Live Database. Retrieved January 4, 2018, from https://myip.ms/browse/blacklist/Blacklist_IP_Blacklist_IP_Addresses_Live_Database_Real-time
[47] FireHol IP Lists. IP Blacklists: IP Reputation Feeds. Retrieved January 6, 2018 from http://iplists.firehol.org/
[48] GreenSnow. Welcome to GreenSnow.co the blacklisted list of IPs for online servers. Retrieved January 6, 2018 from https://greensnow.co/
[49] Opendns.Opendns/public-domain-lists. Retrieved January 6, 2018, from https://github.com/opendns/public-domain-lists
[50] Maxmind. GeoIP2 Precision Insights Service. Retrieved January 6, 2018, from https://www.maxmind.com/en/geoip2-precision-insights
[51] Maxmind. GeoLite2 Free Downloadable Databases. Retrieved January 6, 2018, from https://dev.maxmind.com/geoip/geoip2/geolite2/
[52] Shodan. Retrieved February 14, 2018, from https://www.shodan.io/
[53] Python. 22.28. ipaddress — IPv4/IPv6 manipulation library. Retrieved September 7, 2017, from https://docs.python.org/3/library/ipaddress.html
[54] Spamhaus. The Spamhaus Project. Retrieved September 7, 2017, from https://www.spamhaus.org/
[55] Ines Färber, Stephan Günnemann, Hans-Peter Kriegel, Peer Kröger, Emmanuel Müller, Erich Schubert, Thomas Seidl, and Arthur Zimek. 2010. On using class-labels in evaluation of clusterings. In *MultiClust: 1st international workshop on discovering, summarizing and using multiple clusterings held in conjunction with KDD*, p. 1.

A APPENDICES

A.1 Summary of Network Features

Sno.	Network Feature	Description	Maxmind ASN & City Lite Databases	Maxmind GeoIP2 Precision - Insights (web service)	Shodan	Python Lib	SpamHaus
1	Continent	Continent	-	x	-	-	-
2	Country_Iso_Code	Country's ISO code	x	x	x	-	-
3	Is_In_European_Union	is_in_european_union	-	x	-	-	-
4	Country_Name	Country	x	x	x	-	-
5	Sub_Division_Name	Sub Division/Region	x	x	-	-	-
6	Sub_Division_Iso_Code	-	x	-	x	-	-
7	City_Name	City	x	x	x	-	-
8	Time Zone	Time Zone	-	x	-	-	-
9	Postal_Code	Postal code	x	x	x	-	-
10	Region_Code	Region code	-	-	x	-	-
11	Latitude	Latitude	x	x	x	-	-
12	Longitude	Longitude	x	x	x	-	-
13	Asn	The autonomous system number associated with the IP address	x	x	-	-	-
14	Asn_Org	ASN Organisation	x	x	-	-	-
15	Org	Organisation	-	-	x	-	-
16	Os	Operating system	-	-	x	-	-
17	Area_Code	-	-	-	x	-	-
18	Dma_Code	-	-	-	x	-	-
19	Ports	Port	-	-	x	-	-
20	Hash	-	-	-	x	-	-
21	Transport	Transport protocol	-	-	x	-	-
22	Isp	Internet service provider	-	x	x	-	-
23	Registering Organization	Registering Organization	-	x	-	-	-
24	Anonymous Ip Used?	Anonymous Ip Used?	-	x	-	-	-
25	Proxy Type - Hosting	Proxy type - Hosting	-	x	-	-	-
26	Proxy Type - Public	Proxy type - Public	-	x	-	-	-
27	Tor Exit	Tor exit	-	x	-	-	-
28	Vpn	Virtual private network that encrypts and routes all traffic through the VPN server, including programs and applications.	-	x	-	-	-
29	Asn	The autonomous system number associated with the IP address	-	x	-	-	-
30	Anonymous Ip Type	Anonymous IP Type	-	x	-	-	-
31	Average_Income	Average_Income	-	x	-	-	-
32	User Type	User Type	-	x	-	-	-
33	Population Density	Population Density	-	x	-	-	-
34	User Count	User Count	-	x	-	-	-
35	Domain (Sld/Tld)	Domain (SLD/TLD)	-	x	x	-	-
36	Multicast	Is a multicast IP?	-	-	-	x	-
37	Private	Is a private IP?	-	-	-	x	-
38	Reserved	Is a reserved IP?	-	-	-	x	-
39	Loopback	Is a loopback IP?	-	-	-	x	-
40	Link_Local	Is a link_local IP?	-	-	-	x	-
41	Spam	Any history of being a spam IP?	-	-	-	-	x
		Total Features Per Source	**10**	**26**	**17**	**5**	**1**

A.2 Summary of Geo-contextual Features

Sno.	Dataset Name	Geo-Contextual Feature	Feature Type
1	Gender	mean_gender_value_Both Sexes	Continuous
2	Gender	mean_gender_value_Female	Continuous
3	Gender	mean_gender_value_Male	Continuous
4	Kind of Toilet	mean_toilet_value_Flush/pour flush toilet	Continuous
5	Kind of Toilet	mean_toilet_value_For exclusive use	Continuous
6	Kind of Toilet	mean_toilet_value_For exclusive use - Flush/pour flush toilet	Continuous
7	Kind of Toilet	mean_toilet_value_For exclusive use - Non-flush toilet	Continuous
8	Kind of Toilet	mean_toilet_value_No toilet available	Continuous
9	Kind of Toilet	mean_toilet_value_Non-flush toilet	Continuous
10	Kind of Toilet	mean_toilet_value_Shared	Continuous
11	Kind of Toilet	mean_toilet_value_Shared - Flush/pour flush toilet	Continuous
12	Kind of Toilet	mean_toilet_value_Shared - Non-flush toilet	Continuous
13	Kind of Toilet	mean_toilet_value_Total	Continuous
14	Kind of Toilet	mean_toilet_value_Unknown (whether flush or non-flush inside the housing unit)	Continuous
15	Kind of Toilet	mean_toilet_value_Unknown (whether for exclusive use or shared)	Continuous
16	Kind of Toilet	mean_toilet_value_Unknown (whether for exclusive use toilet is flush/pour flush or not)	Continuous
17	Kind of Toilet	mean_toilet_value_Unknown (whether shared toilet is flush/pour flush or not)	Continuous
18	Kind of Toilet	mean_toilet_value_Unknown (whether toilet is available inside or outside the unit)	Continuous
19	Kind of Toilet	mean_toilet_value_With toilet outside the housing unit	Continuous
20	Kind of Toilet	mean_toilet_value_With toilet within the housing unit	Continuous
21	Kind of Living Quarters	mean_living_quarter_value_Collective living quarters	Continuous
22	Kind of Living Quarters	mean_living_quarter_value_Conventional dwellings	Continuous
23	Kind of Living Quarters	mean_living_quarter_value_Housing units	Continuous
24	Kind of Living Quarters	mean_living_quarter_value_Other housing units	Continuous
25	Kind of Living Quarters	mean_living_quarter_value_Total	Continuous
26	Kind of Living Quarters	mean_living_quarter_value_Unknown	Continuous
27	Kind of Living Quarters	mean_living_quarter_value_Unknown type of housing unit	Continuous
28	Kind of Housing	mean_housing_unit_value_Conventional dwellings	Continuous
29	Kind of Housing	mean_housing_unit_value_Does not have all basic facilities (conventional dwelling)	Continuous
30	Kind of Housing	mean_housing_unit_value_Has all basic facilities (conventional dwelling)	Continuous
31	Kind of Housing	mean_housing_unit_value_Informal housing units	Continuous
32	Kind of Housing	mean_housing_unit_value_Mobile housing units	Continuous
33	Kind of Housing	mean_housing_unit_value_Other housing units	Continuous
34	Kind of Housing	mean_housing_unit_value_Semi-permanent dwellings	Continuous
35	Kind of Housing	mean_housing_unit_value_Total	Continuous
36	Kind of Housing	mean_housing_unit_value_Unknown (type of housing unit)	Continuous
37	Kind of Housing	mean_housing_unit_value_Unknown (type of other housing unit)	Continuous
38	Kind of Housing	mean_housing_unit_value_Unknown (whether or not conventional dwelling has all basic facilities)	Continuous
39	Kind of Fuel used	mean_fuel_type_value_Animal dung	Continuous
40	Kind of Fuel used	mean_fuel_type_value_Charcoal	Continuous
41	Kind of Fuel used	mean_fuel_type_value_Coal	Continuous
42	Kind of Fuel used	mean_fuel_type_value_Crop residues	Continuous
43	Kind of Fuel used	mean_fuel_type_value_Electricity	Continuous
44	Kind of Fuel used	mean_fuel_type_value_Firewood	Continuous
45	Kind of Fuel used	mean_fuel_type_value_Gas	Continuous
46	Kind of Fuel used	mean_fuel_type_value_Kerosene/paraffin (petroleum based)	Continuous
47	Kind of Fuel used	mean_fuel_type_value_Liquified petroleum gas (LPG)	Continuous
48	Kind of Fuel used	mean_fuel_type_value_Oil (ind. vegetable oil)	Continuous
49	Kind of Fuel used	mean_fuel_type_value_Other	Continuous
50	Kind of Fuel used	mean_fuel_type_value_Total	Continuous
51	Kind of Fuel used	mean_fuel_type_value_Unknown	Continuous
52	Kind of Waste Disposal	mean_solid_waste_value_Occupants burn solid waste	Continuous
53	Kind of Waste Disposal	mean_solid_waste_value_Occupants bury solid waste	Continuous
54	Kind of Waste Disposal	mean_solid_waste_value_Occupants compost solid waste	Continuous
55	Kind of Waste Disposal	mean_solid_waste_value_Occupants dispose of solid waste in a local dump not supervised by authorities	Continuous
56	Kind of Waste Disposal	mean_solid_waste_value_Occupants dispose of solid waste in a local dump supervised by authorities	Continuous
57	Kind of Waste Disposal	mean_solid_waste_value_Occupants dispose solid waste into river/sea/creek/pond	Continuous
58	Kind of Waste Disposal	mean_solid_waste_value_Other arrangements	Continuous
59	Kind of Waste Disposal	mean_solid_waste_value_Solid waste collected by self-appointed collectors	Continuous
60	Kind of Waste Disposal	mean_solid_waste_value_Solid waste collected on a regular basis by authorized collectors	Continuous
61	Kind of Waste Disposal	mean_solid_waste_value_Solid waste collected on an irregular basis by authorized collectors	Continuous
62	Kind of Waste Disposal	mean_solid_waste_value_Total	Continuous
63	Kind of Waste Disposal	mean_solid_waste_value_Unknown	Continuous
64	Kind of Water System	mean_water_supply_value_Other	Continuous
65	Kind of Water System	mean_water_supply_value_Piped water inside the unit	Continuous
66	Kind of Water System	mean_water_supply_value_Piped water outside the unit but within 200 metres	Continuous
67	Kind of Water System	mean_water_supply_value_Total	Continuous
68	Kind of Water System	mean_water_supply_value_Unknown	Continuous
69	Communication Infrastructure	mean_ict_type_value_All households	Continuous
70	Communication Infrastructure	mean_ict_type_value_Households accessing the internet from home	Continuous
71	Communication Infrastructure	mean_ict_type_value_Households accessing the internet from other than home	Continuous
72	Communication Infrastructure	mean_ict_type_value_Households having fixed line telephone	Continuous
73	Communication Infrastructure	mean_ict_type_value_Households having mobile cellular telephone	Continuous
74	Communication Infrastructure	mean_ict_type_value_Households having personal computer	Continuous
75	Communication Infrastructure	mean_ict_type_value_Households having radio	Continuous
76	Communication Infrastructure	mean_ict_type_value_Households having television set	Continuous