# Learning Semantic Attributed Graphs for Judging Deviated Human Activity and Understanding

張, 康

# Learning Semantic Attributed Graphs for Judging Deviated Human Activity and Understanding

Kang Zhang

Graduate School of Systems Life Sciences

Kyushu University

A thesis submitted for the degree of

*Doctor of Philosophy*

July 2023

## Abstract

Deviations, which are often viewed as outliers, errors, or noises in data, are prevalent in human activity and understanding. Judging deviated human activity and understanding aims to identify unexpected activities that differ from the normal patterns and biased interpretations of information that can distort the truth and manipulate public opinions. Various tasks and real-world applications have been investigated under this topic, such as abnormal event detection, human monitoring, and fake news detection. Among these tasks, semantic relations are important to understand the entities and their complex associations, such as the semantic consistency between humans and locations in human monitoring and the semantic relevance between sentences in news articles. Since many deviations in human activity and understanding contradict the expected semantic relations, it is crucial to effectively capture and model these relations between entities for identifying the deviations.

In this thesis, we propose learning semantic attributed graphs for two significant tasks within the scope of judging deviated human activity and understanding, i.e., detecting anomalous image regions in human monitoring and judging credible and unethical explanations of statistical data. A semantic attributed graph can provide a structured representation of the complex associations and rich information of entities in the two tasks, such as the regions and their semantic relations in an image as well as the phrases and their semantic similarities in an explanation. Moreover, its explicit modeling of the entities and their relations in the semantic attributed graph enables a development of more accurate detection and judgment algorithms for the two tasks.

We first focus on image region anomaly detection in human monitoring, which aims to identify irregular human behaviors and inappropriate interactions between humans and objects at the region level. Traditional methods typically handle each region separately without taking their associations into consideration. Therefore, these methods cannot detect contextual anomalies which violate regular interactions between humans and objects. Furthermore, prevailing approaches primarily explore visual relations, such as co-occurrence and spatial relations, of regions to model their interactions. However, they neglect the importance of capturing semantic relations among regions, leading to inaccurate predictions of the contextual anomalies. To address their limitations, we introduce a Spatial and Semantic Attributed Graph to represent the regions and their associations in an image. In addition to connecting regions by considering their spatial adjacency, the graph further incorporates semantic relations between re-

gions by leveraging their semantic similarities between their captions. Then we devise a Spatial and Semantic Graph Auto-Encoder (SSGAE) to estimate the abnormality of the regions by jointly reconstructing the attributes and the structures of the proposed graph. Experimental evaluations on three real-world datasets demonstrate that our method outperforms baselines in terms of ROC curves and AUC scores.

Then we focus on judging credible and unethical statistical data explanations which exploit human instincts. We propose that unethical explanations that are credible are more influential and harmful than non-credible ones as they are more likely to be accepted by humans. To judge such explanations, we first devise three phrase embedding-based methods. The conditions in the three methods are designed to compare the semantic relevance between the phrases of subjects and characteristics specified in the explanation. However, experimental results show that counter-intuitive semantic similarities between phrases in the method lead to numerous false predictions. To improve the accuracy of judgment, we introduce a Phrase Similarity Graph to model an explanation by considering more phrases and their semantic similarities. The proposed graph enables generating additional conditions for comparison. Then we devise a credibility score to combine the satisfied conditions and their importance quantified by sub-graph entropy for a more accurate judgment. Our experiments conducted on 14 types of statistical data explanations show the superiority of our proposed method compared with the phrase embedding-based method in terms of accuracy. Scrutiny reveals that our proposed method mitigates the problem of the counter-intuitive semantic similarities at a satisfactory level.

## Acknowledgments

Foremost, I would like to express my gratitude to my esteemed supervisor Professor Einoshin Suzuki for accepting me as his student, for the strict and invaluable academic training to improve my logical thinking, and for his continuous support and patience when I encountered obstacles throughout my doctoral study. His expertise, passion, and dedication to the field of research have been instrumental in shaping my research and developing my skills.

I am deeply grateful to Assistant Professor Tetsu Matsukawa for his academic advice and help with daily affairs in our laboratory. Moreover, I would like to thank the present and past members of our laboratory. My special thanks go to Qiming Zou, Wenbo Li, Ning Dong, Liheng Shen, Jose Alejandro Avellaneda Gonzalez, Hiroaki Shinden, Tatsuki Mutsuro, Muhammad Fikko Fadjrimiratno, Yuanyuan Li, Yusuke Hatae, and Yuichiro Nomura for their assistance and companionships during my doctoral course. Particular acknowledgment is also due to Qiming Zou and Wenbo Li for their insightful discussions and valuable suggestions that enriched my research work. Moreover, I would like to thank JSPS KAKENHI (Grant Number JP21K19795) for supporting a part of the work in the thesis. I also appreciate the financial support provided by the China Scholarship Council (Grant Number 201906330075).

In concluding this journey, I am forever grateful to my parents, whose unwavering support and unconditional love have been the driving force behind my Ph.D. studies. I would also like to extend my heartfelt appreciation to my beloved fiancee for her patience, understanding, and companionship when it was most required. Finally, I would like to thank my dear friends for supporting me whenever I needed them. Their feedback and encouragement have made this challenging journey more meaningful and enjoyable.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Background

Deviations, which can be viewed as outliers, errors, or noise in data [5], widely exist in human activity and understanding. Deviations in human activities refer to human behaviors and interactions between humans and objects that do not conform to the expected or normal patterns [6, 7]. Deviated human activities can be observed in various domains, including surveillance [8], healthcare [9], and public safety [10]. On the other hand, deviations in human understanding refer to the interpretations and explanations of information that deviate from factual and ethical standards [11]. Deviated human understandings can be represented as diverse forms of unethical and biased explanations, such as fake news, rumors, and inflammatory tweets, which can distort human understanding of reality and lead to errors in judgment [12, 13]. Since these deviations pose potential risks to the security and morality of our society, there has been a growing interest in judging deviated human activity [6, 8, 10, 14] and understanding [15–18] in the areas of data mining, machine learning, as well as Artificial Intelligence (AI) and ethics.

In the real world, entities are usually interacting with each other through various relationships [19]. Among these relationships, semantic relations are common and important between entities in human activity and understanding, such as semantic consistency between humans and locations in human monitoring [6] and semantic relevance between phrases in human understanding [18]. However, these semantic relations are not always explicit or easy to identify. A semantic attributed graph [20] is a powerful representation which can explicitly describe the semantic relations between entities, where the nodes are entities with attributes that provide necessary information and the edges are semantic relations that connect them.

Deviations in human activity and understanding often violate expected semantic relations between entities. Specifically, deviated human activities, which are commonly known as anomalies or outliers, usually happen when they contradict their expected relations with scenes or locations [6,8,10]. For example, a man riding on a bicycle down a pedestrian sidewalk is anomalous because the human behavior does not conform to the scene [6]. This kind of anomaly frequently exists in various real-world applications, including abnormal event detection [10] and human monitoring [7, 14]. On the other hand, deviated human understandings, such as unethical and biased explanations can distort the truth and manipulate public opinion often by exploiting human cognitive biases [12] and instincts [2, 18]. To judge such explanations, multiple irregular semantic relations are explored in their words, phrases, or sentences. For example, the relations between sentences or phrases and their polarity or subjectivity are commonly investigated for fake news detection [15]. Consequently, it is necessary to capture and analyze the semantic relations between entities in judging deviated human activity and understanding.

## 1.2 Motivation and Contribution

### 1.2.1 Motivation

In this thesis, we focus on learning semantic attributed graphs for two significant tasks within the problem of judging deviated human activity and understanding, i.e., image region anomaly detection in human monitoring and judging credible and unethical explanations of statistical data.

Understanding and capturing semantic relations between entities is crucial in identifying deviations in the two tasks, including detecting anomalous regions and judging credible and unethical statistical data explanations. In the former task, a context of a region is often characterized by other regions with semantic relations. For example, a kitchen sink and a white chair in the kitchen can be necessary to describe the context of the resting area when detecting anomalies which do not conform to their contexts. However, existing methods have limitations in detecting such anomalies because they either handle each region separately [21–23] or primarily focus on exploring visual relations, such as co-occurrence [24] and spatial relations between regions [7, 25], as the contexts of regions. These methods neglect the importance of relations between regions at the semantic level.

In the latter task, our phrase embedding-based methods employ semantic relations between phrases, including subjects and characteristics, in their conditions for judgment. The conditions are designed to compare the semantic similarities between different kinds of phrases specified in the explanation. However, counter-intuitive semantic similarities between limited phrases of subjects and properties lead to unsatisfactory results on the task.

To overcome these limitations, we propose to learn semantic attributed graphs as

a common approach for the two tasks. The semantic attributed graphs are capable of explicitly representing the semantic relations and attributes of entities, including the regions and their relations in an image and the phrases and their semantic similarities in an explanation, which can be easily managed by detection and judgment algorithms for more accurate identification of the deviations.

### 1.2.2 Contributions of This Thesis

In this thesis, we introduce two main contributions to tackle the two tasks in the problem of judging deviated human activity and understanding.

In Chapter 2, we focus on the image region anomaly detection task. It is a challenging yet important task as it focuses on identifying fine-grained anomalies at the region level [7, 21], including irregular human behaviors and inappropriate interactions between humans and objects. The region anomalies in the task are diverse and complex, including single anomalies and contextual anomalies. A single anomaly refers to an abnormal region which is never observed in normal instances. In addition, a contextual anomaly refers to a region which violates its expected context, where the context of the region is characterized by its interactions and relationships with other regions in the same image. For instance, a man making a phone call is normal in the resting area while abnormal in the working area if the latter is not allowed. Although the two regions depict the same activity, for the purpose of anomaly detection, they had better be distinguished from each other [14]. Therefore, effectively capturing the relations among regions, which represent the contextual information, is critical for the region anomaly detection task.

To effectively detect the diverse region anomalies in human monitoring, we propose a Spatial and Semantic Attributed Graph to model the regions and their contexts

in an image. In contrast to previous methods which primarily consider spatial relations, our spatial and semantic attributed graph further incorporates the relations among regions at the semantic level. The graph leverages both the spatial adjacency among regions and the semantic similarities among their captions to characterize the complex contexts of regions, where a region and its context in an image can be represented as a node and its neighboring nodes, respectively. Then a Spatial and Semantic Graph Auto-Encoder (SSGAE) is devised by adopting a sum aggregation strategy [26] to estimate the abnormality of regions via dual reconstruction optimizations.

In Chapter 3, we focus on the task of judging deviated human understandings of statistical data. Explaining quantitative evidence plays a crucial role in various scientific research methods [27–29]. Specifically, here the deviated human understanding of statistical data refers to an unethical explanation [18], which mainly describes the semantic relevance between the subject and its characteristics based on the statistical data. Among these unethical explanations, we propose those that are credible are more influential and harmful than non-credible ones as they are more likely to be accepted by people. Based on the subjects and characteristics in the explanations, investigating semantic relations between such phrases plays an important role in judging credible and unethical statistical data explanations.

We first define 21 types of such explanations which exploit the human instincts in Rosling et al. [2]. Based on phrase embedding technique, we devise three judgment methods $\alpha$, $\beta$, and $\gamma$ by comparing semantic relevance between phrases specified in the explanation. However, method $\beta$ exhibits low accuracy due to the counter-intuitive semantic similarities between the specified phrases. To address this limitation and achieve better accuracy for judging the credible and unethical statistical data explanations, we propose a graph-based method $\beta^2$. In method $\beta^2$, a Phrase Similarity Graph is constructed to represent each phrase as a node and connects different kinds of phrases

based on their semantic relations. The graph explores semantic similarities between more phrases by considering their synonyms to generate necessary comparison conditions for judgment. Moreover, to quantify the different importance of the generated conditions, we adopt graph entropy to measure the uncertainty of the semantic similarities between nodes in the graph. Lastly, a credibility score is devised by combining the satisfied conditions and their importance for judgment.

## 1.3 Thesis Organization

The rest of this thesis is organized as follows. In Chapter 2, we introduce our graph-based method, including a Spatial and Semantic Attributed Graph and a Spatial and Semantic Graph Auto-Encoder (SSGAE), to tackle the image region anomaly detection task in human monitoring. We evaluate the performance of SSGAE on three real-world datasets compared with several baselines. In Chapter 3, we first define 21 types of credible and unethical explanations of statistical data. Then we introduce three phrase embedding-based methods $\alpha$, $\beta$, and $\gamma$. To improve the low accuracy of method $\beta$, we introduce a new graph-based method $\beta^2$ for more accurate judgment of the credible and unethical statistical data explanations. In Chapter 4, we conclude the thesis and discuss our future work.

# Chapter 2

# Spatial and Semantic Attributed Graph for Region Anomaly Detection in Human Monitoring

## 2.1 Overview

Human monitoring, which focuses on human activities, has drawn attention across various research areas, including video surveillance [8, 30, 31], healthcare [9, 32], and human-computer interaction [33, 34]. Within these areas, anomalies in human activities, e.g., irregular human behaviors and inappropriate interactions between humans and objects, pose a significant problem in many security-related and healthcare scenarios. Such anomalies include abnormal events in video surveillance [10, 35] and unusual signals in medical monitoring [36]. Therefore, anomaly detection in human monitoring, which concentrates on discovering unexpected patterns that deviate from those seen in normal instances, has attracted substantial interest of researchers. It has

a wide range of real-world applications, such as violence detection [37], fall risk discovery [38], and trajectory outlier detection [39].

Among such works, image region anomaly detection [7, 21, 40–43] is a critical task for identifying abnormal areas from images in human monitoring. However, it is challenging to detect region-level anomalies due to their diversity. Traditional methods focus on discovering unobserved regions that deviate from the patterns learned from normal image regions [21, 41–43]. Such a region can be defined as a single anomaly in human monitoring. For instance, the region of a man holding a baseball bat in the laboratory [21] is a single anomaly, as such behavior is never observed in normal regions. In addition to the single anomalies, there also exist contextual anomalies [7, 40], which involve violations of regular interactions among humans and objects, as the context of a region is characterized by other regions in the same image. For instance, the region of a man making a phone call is normal when it is located close to a kitchen sink and a soap bottle in an image, as they are in a resting area, while abnormal when close to a bookshelf and a notebook PC in another image, as they are in a working area if the latter is not allowed. Figure 2.1 shows several examples of single anomalies and contextual anomalies compared with normal regions. The single anomalies, e.g., a man holding a baseball bat and a man holding an umbrella, have significant visual differences from the normal regions. In contrast, contextual anomalies may exhibit similar human behaviors to those in normal instances. Therefore, capturing contextual information is crucial in the region anomaly detection task.

As mentioned above, several region-level anomaly detection methods typically concentrate on identifying single anomalies without considering the contexts of regions [21, 41–43]. These methods mainly discover patch-level deviations by learning the regularities of normal instances. On the other hand, recent approaches have been proposed for detecting both single and contextual anomalies by exploring the relation-

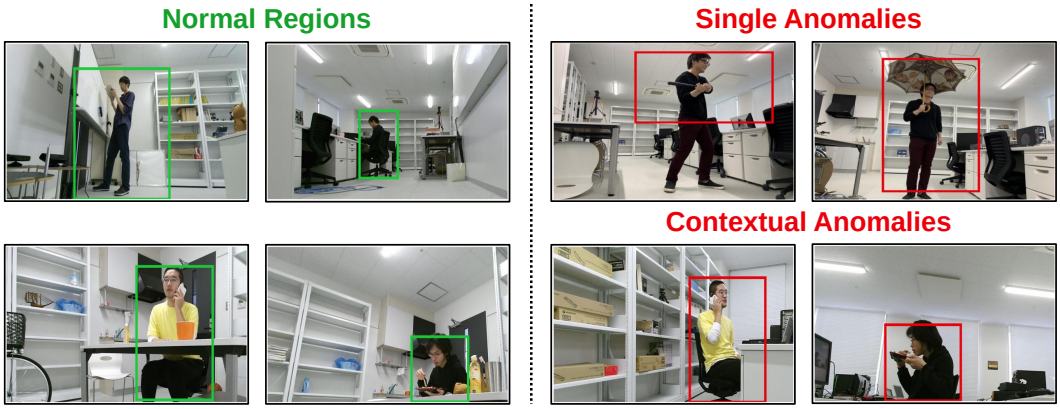



Figure 2.1: Examples of single anomalies and contextual anomalies compared with normal regions. Green rectangles indicate normal regions, while red rectangles indicate region anomalies. The first row shows single anomalies, including a man holding a baseball bat and an umbrella, which are not observed in normal regions. The second row shows contextual anomalies, including a man making a phone call and eating, which are not allowed in a working area. All the regions are generated by a deep-captioning model Densecap [1].

ships among regions as their contexts. They can be classified into an object-label-based method [24], a spatial-relation-based method [25], and a deep-captioning-based method [7]. Choi et al. [24] represent all the objects in an image with a tree-structured model to detect objects that do not conform to the scene. However, utilizing all object labels beforehand is impractical for the anomaly detection task. The spatial-relation-based method [25] considers the positions, such as above, below, and inside, of two objects to detect abnormal semantic relationships between a pair of image segmentations, while such spatial positions are limited in characterizing diverse region contexts which are essential for detecting the contextual anomalies. In addition to exploiting visual features of image regions, deep-captioning-based methods [7, 21] adopt deep-captioning models, such as DenseCap [1], to obtain region captions as the semantic information for the target task. Since these methods also consider both the visual and semantic information of image regions on the same task, they are the most relevant

works to our proposed method. They focus on detecting anomalous single regions and anomalous region pairs by exploring the spatial relations between two regions. Nevertheless, they do not consider interactions among more than two regions and are thus limited in capturing the complex context for detecting contextual anomalies in human monitoring.

To capture the contexts of image regions for more accurate region anomaly detection in human monitoring, we propose using graph structures to model regions and their relations in an image. Graphs are capable of representing complex relations among data points, making their interactions explicit and easily manageable by graph-based algorithms. The superiority of graph structures has been demonstrated in many visual tasks, including representing the co-occurrence of regions in an image for object detection [44] and the spatio-temporal relations of regions in video clips for action recognition [45]. In addition, several existing methods for frame-level video anomaly detection propose to model the spatio-temporal relations of regions by graphs for better performance. For example, A spatio-temporal context graph [46] is constructed to model visual context information including appearances of objects, as well as spatio-temporal relationships among objects for discriminating abnormal events. Considering the spatial similarity and the temporal consistency in video data, a spatio-temporal graph-based deep model [47] is devised for detecting frame-level anomalies. However, simply considering the co-occurrence cannot effectively capture the complex relations of image regions. For instance, some regions should be more strongly correlated, e.g., a man and an umbrella in the rightmost image in Figure 2.1, compared with others due to their close spatial relations. On the other hand, temporal patterns do not exist in image regions within a single image.

In this Chapter, to address the aforementioned limitations, we introduce a graph-based framework, including a Spatial and Semantic Attributed Graph and a Spatial and

Semantic Graph Auto-Encoder (SSGAE), to tackle the region anomaly detection task. In our method, the Spatial and Semantic Attributed Graph is proposed to model regions and their contexts within an image by leveraging the spatial and semantic relations among regions. Specifically, the graph provides a structured representation of each region with features as a node with attributes[1] and connect them by considering the spatial adjacency among regions and the semantic similarities among their captions.

Then a tailored graph auto-encoder, SSGAE, is devised for detecting anomalous nodes in the graph via dual reconstruction tasks. In particular, since the regions depicting similar objects, such as a desk, and similar human behaviors, such as a man sitting on a chair, frequently appear in human monitoring, the neighbors of a node usually contain similar features in the graph. The mean-pooling or max-pooling strategy captures the proportions of the node attributes or the most representative node attribute as the representation of node neighbors. Therefore, existing graph auto-encoders [48, 49] equipped with these strategies are difficult to discriminate such node neighbors representing the regional contexts. Consequently, SSGAE adopts the sum aggregation strategy used in Graph Isomorphism Network (GIN) [26], which is superior in discriminating such node neighbors by capturing all their attributes, as we will give the details in Chapter 2.4.3.1.

In summary, the contributions of this Chapter are as follows:

- We propose a graph-based framework, including a Spatial and Semantic Attributed Graph and SSGAE, to tackle the region anomaly detection task. By leveraging the visual and semantic information, the Spatial and Semantic Attributed Graph characterizes the regions with their contexts based on the spatial and semantic relations among the regions and their captions, respectively.

[1] Node attributes and node features are utilized interchangeably in this Chapter.

- For more accurate region anomaly detection, we devise a customized graph auto-encoder, SSGAE. SSGAE adopts a sum aggregation strategy [26] to effectively capture the structure information and detect anomalous nodes in the graph by jointly reconstructing the node features and structures in the Spatial and Semantic Attributed Graph.

- We conduct extensive experiments on three real-world datasets to evaluate the performance of our method. The experimental results show that SSGAE outperforms other advanced anomaly detection methods, which demonstrates the effectiveness of SSGAE on the region anomaly detection task.

## 2.2 Related Work

In this section, we mainly introduce related works on two topics: (1) image and region anomaly detection and (2) graph anomaly detection.

### 2.2.1 Image and Region Anomaly Detection

Image-level and region-level anomaly detection have been active research topics for decades, which can be classified into two categories: those which implicitly consider the relationships among images or regions and those which explicitly consider them. The former methods mainly focus on discovering pixel-wise or patch-level deviations by learning regularities of normal instances, such as defect detection [42, 50] and medical image analysis [22, 23]. These works have shown their advantages in detecting anomalous regions via self-supervised learning [41,42,51,52], where the contextual information characterized by other regions is implicit in their tasks. Since these methods consider images or regions separately, they are unable to detect contextual anomalies

in human monitoring.

On the other hand, the latter methods explicitly combine the images or regions with their relationships as the contexts to understand and discover diverse image-level or region-level anomalies, such as video surveillance [10, 35] and human monitoring [7, 14, 21]. Among such works, several approaches [24, 25, 40, 46] consider the regions and their relations in the visual perspective for region anomaly detection, while our previous methods [7,21] additionally adopt deep-captioning models, such as DenseCap [1], to obtain region captions as the semantic information for the task. Sun et al. [46] proposed a Spatio-Temporal Graph (STG) to represent spatio-temporal relations among objects to bridge the gap between an anomaly and its context. Moreover, Spatial-Temporal Graph-based Convolutional Neural Networks (STGCNs) [47] construct a spatial similarity graph and a temporal consistency graph with a self-attention mechanism to model the correlations of video clips for video anomaly detection. Choi et al. [24] identified out-of-context objects, i.e., objects which do not conform to the scene, by modeling all the objects in the same image via a tree-based graphical model. These works have shown the effectiveness of utilizing graphical models to represent the relationships among video clips or objects for video or region anomaly detection. To detect anomalous images in human monitoring, Dong et al. [40] employed inpainting techniques to coarsen image regions and then generate the regions by utilizing the remaining part of the image. Moreover, Semantic Anomaly Detection (SAD) [25] models the relative positions and sizes of all object pairs to detect abnormal semantic relationships between a pair of image segmentations. These methods have proven their superiority in exploring visual information of videos and images to detect abnormal instances. However, in addition to the visual features and relations of image regions considered by these methods, region captions provide semantic information regardless of intra-object variations, which can contribute to more accurate region anomaly

detection [7, 21]. Our previous methods [7, 21] exploit both the visual features of regions and the semantic information of region captions for the target task. Nevertheless, they consider each region separately for the anomalous single regions [21] as well as the relations of two overlapping regions for anomalous region pairs [7]. Therefore, they cannot capture the relations among more than two regions that indicate the region context, leading to failures in detecting some of the contextual anomalies in our task.

### 2.2.2 Graph Anomaly Detection

Graph Neural Networks (GNNs), which are a family of deep learning models for graph or node embedding [53], have been widely explored for graph anomaly detection. Graph contrastive learning methods [54–56] sample well-designed instance pairs, which consist of nodes and their neighboring structures, to devise contrastive learning models for graph anomaly detection. However, to achieve a satisfactory performance, elaborate handcrafted contrastive pretext tasks are mandatory for such kind of methods. On the other hand, several reconstruction-based graph auto-encoder frameworks with different neighborhood aggregation strategies are devised for the task. Deep Anomaly Detection on Attributed Networks (DOMINANT) [49] constructs a graph auto-encoder model equipped with Graph Convolutional Network (GCN) [48] layers to reconstruct the node attributes and structures for detecting abnormal nodes on large-scale graphs. Furthermore, Anomaly Dual Auto-Encoders (AnomalyDAE) [57] tackle the same problem via reconstruction by designing a dual auto-encoder with graph attention layers [58]. By adopting graph attention layers in both the encoder and the decoder, Graph Attention Auto-Encoder (GATE) [59] exhibits a superior performance in learning node representations for node classification.

The existing graph auto-encoders are effective for learning typical node represen-

tations for downstream tasks, such as graph anomaly detection [49, 57] and node classification [59]. However, the learned representations do not explicitly consider all the features in node neighbors since they focus on capturing the proportions of the features or the most representative feature in node neighbors [26]. This limitation would cause failures in discriminating the representations of different node neighbors, which indicates the contextual information of regions would be useful for detecting anomalies in human monitoring.

## 2.3 Problem Formulation

In this Chapter, we utilize bold lowercase Roman letters (e.g., $\mathbf{x}$), bold uppercase Roman letters (e.g., $\mathbf{X}$), and uppercase calligraphic fonts (e.g., $\mathcal{D}$) to denote vectors, matrices, and sets, respectively. All important notations are summarized in Tables 2.1 and 2.2 for convenience.

Our target problem is to detect anomalous regions in human monitoring images [7, 21]. Due to the diversity and rareness of anomalies, anomaly detection is typically solved under a one-class anomaly detection scenario, which means that only normal data is accessible during the training stage [7, 14, 21, 60]. We follow this paradigm in our method. In the target problem, the input dataset $\mathcal{D}$ is composed of training set $\mathcal{D}^{\text{train}} = \{I^k | k = 1, \ldots, K\}$ and test set $\mathcal{D}^{\text{test}} = \{I^{k'} | k' = 1, \ldots, K'\}$, where $I^k$ and $I^{k'}$ denote the images in the training and test sets, respectively. In the training phase, each input image $I^k$ contains a number of $n$ salient regions $r_i^k$ with captions $c_i^k$ and region labels $y_i^k$ as $\{(r_i^k, c_i^k, y_i^k) | i = 1, \ldots, n\}$. Since we tackle the target problem in the one-class anomaly detection scenario, $\mathcal{D}^{\text{train}}$ only contains normal regions, in which $y_i^k = 0$ denotes the class label of the normal region. In the test phase, each image $I^{k'}$ contains $n$ salient regions with captions and region labels $y_i^{k'} \in \{0, 1\}$ as

Table 2.1: Notations of variables for Spatial and Semantic Attributed Graph.

| Notation | Description |
|---|---|
| $I^k$ | The $k^{th}$ image |
| $r_i^k$ | The $i^{th}$ region in the $k^{th}$ image $I^k$ |
| $c_i^k$ | The caption of the $i^{th}$ region $r_i^k$ |
| $\mathbf{r}_i^k \in \mathbb{R}^{d_r}$ | The visual feature vector of the $i^{th}$ region $r_i^k$ |
| $\mathbf{c}_i^k \in \mathbb{R}^{d_c}$ | The semantic feature vector of caption $c_i^k$ of the $i^{th}$ region |
| $\mathcal{G}^k = \{\mathbf{A}^k, \mathbf{X}^k\}$ | The attributed graph for image $I^k$ |
| $v_i^k$ | The $i^{th}$ node in the graph $\mathcal{G}^k$ |
| $\mathcal{N}(v_i^k)$ | The set of the neighbors adjacent to node $v_i^k$ |
| $\mathbf{A}^k \in \mathbb{R}^{n \times n}$ | The adjacency matrix of graph $\mathcal{G}^k$ |
| $\mathbf{a}_i^k \in \mathbb{R}^d$ | The edge, i.e., structure, information of node $v_i^k$ in $\mathbf{A}^k$ |
| $\mathbf{X}^k \in \mathbb{R}^{n \times d}$ | The node attribute matrix of graph $\mathcal{G}^k$ |
| $\mathbf{x}_i^k \in \mathbb{R}^d$ | The $i^{th}$ node feature vector of node $v_i^k$ |
| $n$ | The number of regions in image $I^k$ and nodes in graph $\mathcal{G}^k$ |
| $d$ | The dimension of node feature |
| $d_r, d_c$ | The dimensions of the visual feature and the semantic feature |

$\{(r_i^{k'}, c_i^{k'}, y_i^{k'}) | i = 1, \ldots, n\}$, where $y_i^{k'} = 1$ denotes the class label of the abnormal region. The output of the target problem is the degree of abnormality for each region $r_i^{k'}$ in $I^{k'}$ from $\mathcal{D}^{\text{test}}$.

Following previous methods for the anomaly detection task [7,14,21,61], we adopt the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) score as the evaluation metric to quantify the performance of our method. The ROC curve is plotted by the true positive rate (TPR) and the false positive rate (FPR)

Table 2.2: Notations of variables and parameters for SSGAE.

| Notation | Description |
|---|---|
| $\mathbf{H}^{(l)} \in \mathbb{R}^{n \times d_l}$ | The hidden representation matrix of graph $\mathcal{G}^k$ in the $l^{th}$ layer of the attributed graph encoder |
| $\mathbf{h}_i^{(l)} \in \mathbb{R}^{d_l}$ | The hidden representation vector of node $v_i^k$ in $\mathbf{H}^{(l)}$ |
| $\mathbf{Z}^k \in \mathbb{R}^{n \times d_e}$ | The final hidden embedding matrix of nodes in graph $\mathcal{G}^k$ |
| $\mathbf{z}_i^k \in \mathbb{R}^{d_e}$ | The final hidden embedding vector of node $v_i^k$ |
| $\widehat{\mathbf{H}}^{(l)} \in \mathbb{R}^{n \times d_l}$ | The hidden representation matrix of graph $\mathcal{G}^k$ in the $l^{th}$ layer of the graph attribute decoder |
| $\widehat{\mathbf{h}}_i^{(l)} \in \mathbb{R}^{d_l}$ | The hidden representation vector of node $v_i^k$ in $\widehat{\mathbf{H}}^{(l)}$ |
| $\mathbf{\Theta}^{(l)} \in \mathbb{R}^n$ | The learnable parameter vector in the $l^{th}$ layer |
| $\mathbf{\Theta}_i^{(l)}$ | The $i^{th}$ learnable parameter in $\mathbf{\Theta}^{(l)}$ |
| $\text{MLP}_{\text{Enc}}^{(l)}, \text{MLP}_{\text{Att}-\text{Dec}}^{(l)}$ | The multi-layer perception modules in the $l^{th}$ layer of the attributed graph encoder and the graph attribute decoder |
| $\text{MLP}_{\text{Str}-\text{Dec}}$ | The multi-layer perception module in the graph structure decoder |
| $L$ | The number of the hidden layers |
| $\beta$ | The hyper-parameter to balance the attribute and the structure reconstruction errors in the objective function |
| $d_l, d_e$ | The dimensions of hidden representation $\mathbf{h}_i^{(l)}$ and final hidden embedding $\mathbf{z}_i^k$ |
| $\widehat{\mathbf{X}}^k, \widehat{\mathbf{A}}^k$ | The reconstructions of $\mathbf{X}^k$ and $\mathbf{A}^k$ |
| $\widehat{\mathbf{x}}_i^k, \widehat{\mathbf{a}}_i^k$ | The reconstructions of $\mathbf{x}_i^k$ and $\mathbf{a}_i^k$ for node $v_i^k$ |
| $s_i^{k'}$ | The anomaly score of node $v_i^{k'}$ in the test phase |

with a range of thresholds. AUC score stands for the value of the area under the ROC curve, which corresponds to the probability that a positive test sample is ranked higher than a negative test sample.

## 2.4 Methodology

### 2.4.1 Pipeline of Our Method

The overall pipeline of our method is shown in Figure 2.2. Our method consists of two modules, including the Spatial and Semantic Attributed Graph and the Spatial and the Semantic Graph Auto-Encoder, which we will introduce in Chapter 2.4.2 and Chapter 2.4.3, respectively. Given an image containing regions with captions, we first construct the Spatial and Semantic Attributed Graph to represent the regions and their spatial and semantic relations, which transforms the region anomaly detection into a graph anomaly detection task. Subsequently, we introduce a customized graph auto-encoder, SSGAE, to tackle the transformed task through dual reconstruction optimizations.

### 2.4.2 Spatial and Semantic Attributed Graph

In both the training and test phases, we generate the regions with captions from images and extract their visual and semantic features through pre-trained deep models [21]. Based on the generated regions with their extracted features, we introduce the criteria for constructing the graph for each image to represent regions with their spatial and semantic relations.

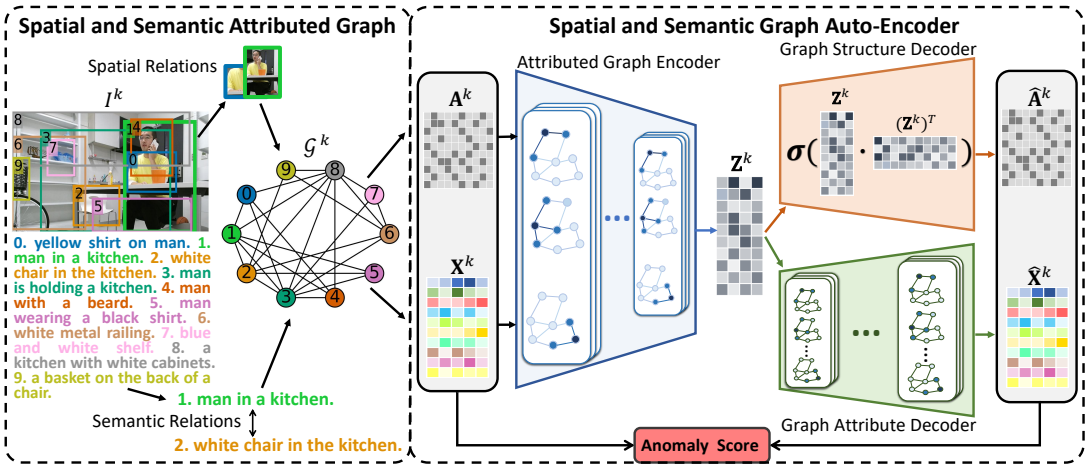



Figure 2.2: The overall pipeline of our method. The left module is the construction of the Spatial and Semantic Attributed Graph. The right module is the architecture of the SSGAE.

#### 2.4.2.1 Generating Regions and Captions from an Image

Following our previous works [7, 14, 21], we adopt a dense captioning model DenseCap [1] to generate region candidates with captions from image $I^k$ and select the top-$n$ salient regions $\{r_i^k | i = 1, \ldots, n\}$ with captions $\{c_i^k | i = 1, \ldots, n\}$ from the region candidates. An example of an image containing the generated regions with captions is shown in the left part of Figure 2.3. Then we explore the visual and semantic information of regions from image $I^k$. Specifically, we utilize an image classification model, ResNet [62], and a sentence embedding model, SBERT [63], to extract visual features of regions $\{\mathbf{r}_i^k | i = 1, \ldots, n\}$ and semantic features of their captions $\{\mathbf{c}_i^k | i = 1, \ldots, n\}$, respectively.

#### 2.4.2.2 Construction of a Spatial and Semantic Attributed Graph

In human monitoring, humans and objects often appear with specific spatial relations to one another in an image. For example, a human, a computer screen, and a desk

typically appear in a regular arrangement [24]. Moreover, the region captions indicate their relations at the semantic level. For example, the two region captions: "man in a kitchen" and "white chair in the kitchen", are highly related to each other to characterize the resting area. Consequently, exploring such spatial and semantic relations among regions is promising to represent their contexts.

We propose the Spatial and Semantic Attributed Graph $\mathcal{G}^k$ to model regions $\{r_i^k | i = 1, ..., n\}$ with their spatial and semantic relations in image $I^k$. Following previous works on graph anomaly detection [49,54,56], we define an attributed graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, where $\mathcal{V} = \{v_1, \ldots, v_n\}$ represents the set of nodes ($|\mathcal{V}| = n$) and $\mathcal{E}$ represents the set of edges ($|\mathcal{E}| = m$). $\mathbf{X} \in \mathbb{R}^{n \times d}$ represents the node attribute matrix, where vector $\mathbf{x}_i \in \mathbb{R}^d$ in $\mathbf{X}$ in the $i^{th}$ row denotes the attribute of the $i^{th}$ node with the dimension $d$. The topology of $\mathcal{G}$ can be denoted by adjacency matrix $\mathbf{A}$, where $\mathbf{A}_{ij} = 1$ represents that there exists an edge between nodes $v_i$ and $v_j$, otherwise $\mathbf{A}_{ij} = 0$. The row vector $\mathbf{a}_i \in \mathbb{R}^n$ in $\mathbf{A}$ denotes the edge information, i.e., the structure, of the $i^{th}$ node. Therefore, the attributed graph can also be denoted as $\mathcal{G} = (\mathbf{A}, \mathbf{X})$.

In the Spatial and Semantic Attributed Graph, a node, its attribute, and its structure information represent a region, its feature, and its spatial and semantic relations with other regions in an image, respectively. Formally, in image $I^k$, region $r_i^k$ is represented as node $v_i^k$ in graph $\mathcal{G}^k$. As mentioned in Chapter 2.2.1, region captions can provide semantic information regardless of intra-object variations for more accurate anomaly detection [21]. Therefore, we concatenate the visual feature of the region and the semantic feature of its caption Concat$(\mathbf{r}_i^k, \mathbf{c}_i^k)$ as the node attribute $\mathbf{x}_i^k$.

Moreover, we make the assumption that the spatially adjacent regions and the regions whose captions have high semantic similarities are informative to characterize the contextual information. To capture these relations, we build spatial edges between nodes when their corresponding regions spatially overlap each other and seman-

tic edges when their region captions have high semantic similarities. Following the previous works on semantic textual tasks [63–65], we utilize cosine similarity as the semantic similarity $\mathrm{Sim}(\cdot)$ between the region captions, which is computed as follows.

$$\mathrm{Sim}(c_i^k, c_j^k) = \frac{\mathbf{c}_i^k \cdot \mathbf{c}_j^k}{\|\mathbf{c}_i^k\| \|\mathbf{c}_j^k\|}. \tag{2.1}$$

If $\mathrm{Sim}(c_i^k, c_j^k) > \theta_{\mathrm{sim}}$, where $\theta_{\mathrm{sim}}$ is a similarity threshold, the two region captions $c_i^k$ and $c_j^k$ are judged to have high semantic relations, and thus a semantic edge is built between nodes $v_i^k$ and $v_j^k$. By building the spatial and semantic edges, the structure information of node $v_i^k$ can be represented as $\mathbf{a}_i^k$, which represents the contexts of region $r_i^k$ in image $I^k$. In this setting, the training set $\mathcal{D}^{\mathrm{train}}$ and test set $\mathcal{D}^{\mathrm{test}}$ can be represented as $\mathcal{G}_{\mathrm{train}}^k = \{\mathbf{A}^k, \mathbf{X}^k\}_{k=1}^K$ and $\mathcal{G}_{\mathrm{test}}^{k'} = \{\mathbf{A}^{k'}, \mathbf{X}^{k'}\}_{k'=1}^{K'}$, respectively.

Figure 2.3 shows an example of constructing a Spatial and Semantic Attributed Graph to model regions in an image. The no. 1 region with its features is represented as node 1 with its attribute, respectively. The edges between nodes 1 and 0, as well as nodes 1 and 2, are built according to their spatially adjacent regions and the high semantic similarities of their captions, respectively.

### 2.4.3 Spatial and Semantic Graph Auto-Encoder

We propose SSGAE to tackle the target problem by detecting abnormal nodes in the Spatial and Semantic Attributed Graph. The architecture of SSGAE is shown in the right module of our proposed method in Figure 2.2. We present the overall procedure of SSGAE, including the training and test phrases, in Algorithm 2 in Appendix A.

With a graph auto-encoder [66] as a backbone, SSGAE consists of three components: an attributed graph encoder, a graph structure decoder, and a graph attribute

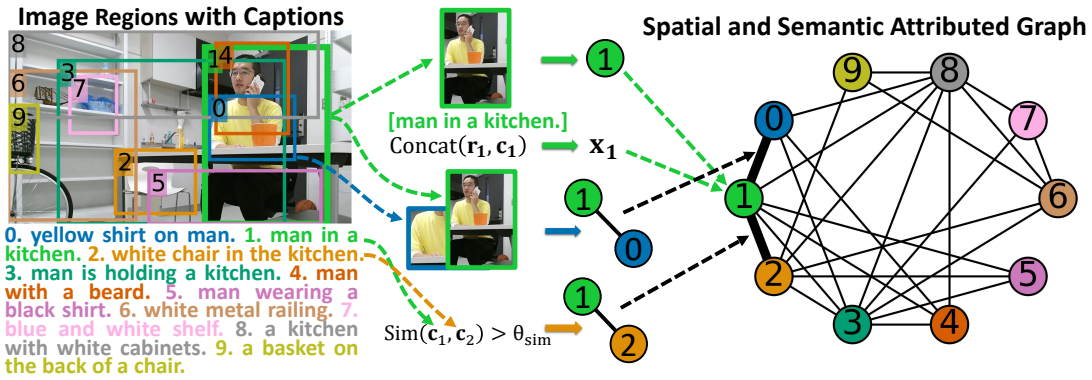



Figure 2.3: Example of constructing a Spatial and Semantic Attributed Graph to model regions in an image. The numbers and colors of the regions in the image and the nodes in the graph correspond to each other.

decoder. Given the constructed graphs as input, SSGAE is devised to estimate the abnormality of each node in each graph by leveraging the node structure and the node attribute reconstruction errors. In particular, we adopt the sum aggregation strategy from GIN [26] in SSGAE to discriminate the diverse node neighbors containing similar node features in the constructed graphs. We will explain the details in Chapter 2.4.3.1.

#### 2.4.3.1 Sum Neighborhood Aggregation Strategy

Different from prevalent graph auto-encoder variants [49, 57, 59, 66], SSGAE adopts the sum neighborhood aggregation strategy from GIN [26]. The mean-pooling or max-pooling aggregation strategies in graph auto-encoders [49, 66] are capable of capturing the proportions of features or the representative feature in node neighbors, respectively. They have shown their advantages in graph anomaly detection on citation networks and social networks, in which the node features are diverse and rarely identical, as the proportions of features or the representative feature in node neighbors already provide strong signals for the task. However, in human monitoring, it is common to have

regions depicting similar objects, such as a desk, and similar human behaviors, such as a man sitting on a chair, appearing frequently in images. This leads to the situation that similar node features often exist in the neighbors of a node in the constructed graphs. In such a case, the sum neighborhood aggregation strategy [26] is capable of explicitly capturing all the features in node neighbors compared with mean-pooling, max-pooling, and weighted average via attention strategies[1] [57, 59].

Figure 2.4 illustrates toy examples to show the advantage of the sum aggregation strategy in discriminating such node neighbors. The no. 0 regions in $I^i$ and $I^j$ and their corresponding nodes are abnormal and normal in red and green colors, respectively. We assume the features of the regions in orange showing laboratory furniture are similar, and the features of the regions in blue showing the black pants are similar. We observe that the mean-pooling or max-pooling strategies aggregate the two kinds of node neighbors into approximately equivalent representations and thus cannot discriminate them well. In contrast, the sum strategy compresses the two kinds of node neighbors into discriminative representations. Consequently, we adopt the sum aggregation strategy in SSGAE since discriminating the representations of such node neighbors, which represent the context of regions, plays a critical role in the region anomaly detection task.

#### 2.4.3.2 Attributed Graph Encoder

To learn discriminative embeddings from the node attributes and structures, the hidden layers in the attributed graph encoder are equipped with the sum aggregation strategy [26] to compress node representations in the aggregation and transformation scheme. Formally, given the constructed spatial graph $\mathcal{G}^k = \{\mathbf{A}^k, \mathbf{X}^k\}_{k=1}^K$, the node

[1] The weighted average via attention strategy may implicitly capture all the node features by learning different weights for node neighbors.

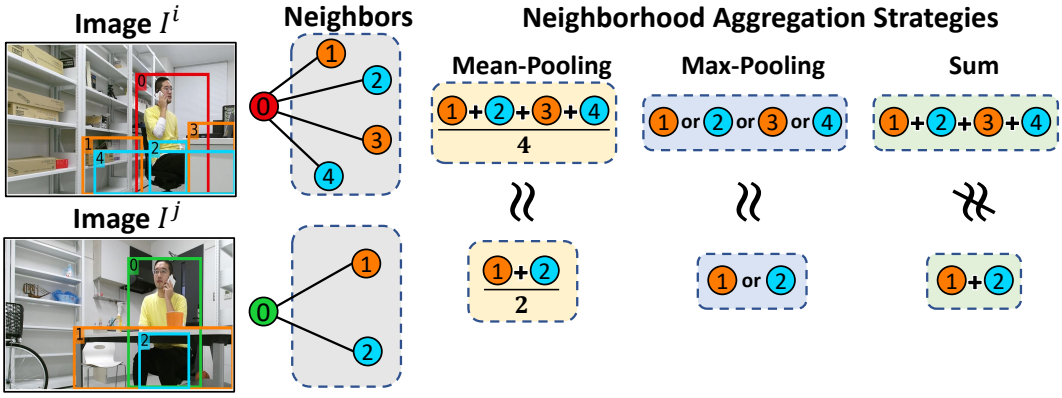



Figure 2.4: Toy examples for different aggregation strategies to discriminate the neighbors of the no. 0 regions in $I^i$ and $I^j$. The numbers and colors of regions in the image and the nodes in the graph correspond to each other.

representation $\mathbf{h}_i^{(l)}$ in the $l^{th}$ layer is iteratively updated as

$$\mathbf{h}_i^{(l)} = \mathrm{MLP}_{\mathrm{Enc}}^{(l)}\left(\left(1+\Theta_i^{(l)}\right)\mathbf{h}_i^{(l-1)} + \sum\nolimits_{v_j^k \in \mathcal{N}(v_i^k)} \mathbf{h}_j^{(l-1)}\right), \tag{2.2}$$

where $\mathrm{MLP}_{\mathrm{Enc}}^{(l)}$ represents the multi-layer perceptron module which adopts the $\mathrm{ReLU}(\cdot)$ activation function [67] in the $l^{th}$ hidden layer of the encoder. We initialize $\mathbf{h}_i^{(0)} = \mathbf{x}_i^k$ as the feature of node $v_i^k$. In the view of the whole matrix, the hidden representation matrix $\mathbf{H}^{(l)}$ is formulated as

$$\mathbf{H}^{(l)} = \mathrm{MLP}_{\mathrm{Enc}}^{(l)}\left(\left(\mathbf{A}^k + \left(1+\mathbf{\Theta}^{(l)}\right)\cdot\mathbf{I}\right)\cdot\mathbf{H}^{(l-1)}\right). \tag{2.3}$$

Here $\mathbf{H}^{(0)} = \mathbf{X}^k$ is the input node attribute matrix. After applying this procedure to $L$ hidden layers, the final hidden embedding matrix is generated as $\mathbf{H}^{(L)} = \mathbf{Z}^k$, where $\mathbf{Z}^k$ is composed of embedding $\mathbf{z}_i^k$ of each node $v_i^k$ in $\mathcal{G}^k$.

#### 2.4.3.3 Graph Structure Decoder

The node structure information, which is represented as the edges of the node connecting other nodes, indicates the contexts of the region. To learn the contextual information for detecting anomalous regions, the graph structure decoder is devised by reconstructing the structure information of nodes. With the final hidden embedding matrix $\mathbf{Z}^k$ as input, the graph structure encoder utilizes the inner product operation, which has been widely employed by [49, 57, 66], with an additional MLP module $\text{MLP}_{\text{Str}-\text{Dec}}$ to estimate the probability of edge $\widehat{\mathbf{A}}_{ij}^k$ between nodes $v_i^k$ and $v_j^k$ as

$$P\left(\widehat{\mathbf{A}}_{ij}^k | \mathbf{z}_i^k, \mathbf{z}_j^k\right) = \sigma\left(\text{MLP}_{\text{Str}-\text{Dec}}\left(\mathbf{z}_i^k \cdot \mathbf{z}_j^{kT}\right)\right), \tag{2.4}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function and $\text{MLP}_{\text{Str}-\text{Dec}}$ adopts the $\text{ReLU}(\cdot)$ activation function. The total reconstructed adjacency matrix $\mathbf{A}^k$ of $\mathcal{G}^k$ is calculated as

$$\widehat{\mathbf{A}}^k = \sigma\left(\text{MLP}_{\text{Str}-\text{Dec}}\left(\mathbf{Z}^k \cdot \mathbf{Z}^{k\text{T}}\right)\right). \tag{2.5}$$

#### 2.4.3.4 Graph Attribute Decoder

The node attribute is composed of both the visual and semantic features extracted from the corresponding region, providing valuable information for its content. To capture the node attribute information, the graph attribute decoder is devised to decompress $\mathbf{Z}^k$ for reconstructing the original node attributes. Similar to the attributed graph encoder, we utilize the same hidden layers by adopting the sum aggregation strategy. The node representation $\widehat{\mathbf{h}}_i^{(l)}$ in the $l^{th}$ layer is computed as

$$\widehat{\mathbf{h}}_i^{(l)} = \text{MLP}_{\text{Att}-\text{Dec}}^{(l)}\left(\left(1+\Theta_i^{(l)}\right)\widehat{\mathbf{h}}_i^{(l-1)} + \sum\nolimits_{v_j^k \in \mathcal{N}(v_i^k)} \widehat{\mathbf{h}}_j^{(l-1)}\right). \tag{2.6}$$

The multi-layer perceptron module $\text{MLP}^{(l)}_{\text{Att}-\text{Dec}}$ in the graph attribute decoder also adopts the $\text{ReLU}(\cdot)$ activation function, where the fully-connected layers are symmetric to the hidden layers in $\text{MLP}^{(l)}_{\text{Enc}}$ in terms of the number of their hidden units for reconstruction. Accordingly, the total hidden representation matrix $\widehat{\mathbf{H}}^{(l)}$ is computed as

$$\widehat{\mathbf{H}}^{(l)} = \text{MLP}^{(l)}_{\text{Att}-\text{Dec}}\left(\left(\mathbf{A}^k + \left(1+\boldsymbol{\Theta}^{(l)}\right)\cdot\mathbf{I}\right)\cdot\widehat{\mathbf{H}}^{(l-1)}\right). \tag{2.7}$$

The input to the graph attribute decoder is $\widehat{\mathbf{H}}^{(0)} = \mathbf{Z}^k$, and the output in the $L^{th}$ layer is the reconstructed node attribute matrix $\mathbf{H}^{(L)} = \widehat{\mathbf{X}}^k$.

#### 2.4.3.5 Objective Function

As suggested in the typical graph auto-encoders [49, 57], the disparities between the node attribute and its reconstruction, as well as the node structure and its reconstruction, provide strong signals to estimate the abnormality of the node. Following this assumption, we optimize SSGAE by jointly minimizing structure reconstruction error $\mathcal{L}_{\text{str}}$ and attribute reconstruction error $\mathcal{L}_{\text{att}}$. Formally, objective function $\mathcal{L}$ of SSGAE is formulated as

$$\mathcal{L} = (1-\beta)\,\mathcal{L}_{\text{str}} + \beta\mathcal{L}_{\text{att}} \tag{2.8}$$

$$= \frac{1}{K}\sum_{k=1}^{K}\left((1-\beta)\,\|\widehat{\mathbf{A}}^k - \mathbf{A}^k\|_F^2 + \beta\|\widehat{\mathbf{X}}^k - \mathbf{X}^k\|_F^2\right), \tag{2.9}$$

where $\beta$ is a hyper-parameter to balance $\mathcal{L}_{\text{str}}$ and $\mathcal{L}_{\text{att}}$ and $\|\cdot\|_F$ denotes the Frobenius norm.

#### 2.4.3.6 Anomaly Score

As we mentioned in Chapter 2.3, the target problem is tackled in a one-class anomaly detection scenario in which only normal data are available in the training stage. Trained on graphs which contain only normal nodes, SSGAE is capable of reconstructing high-quality attributes and structures of the normal nodes [49] by optimizing the objective function. Therefore, in the test stage, the model is supposed to output a high attribute reconstruction error or a high structure reconstruction error for an abnormal node in the test set. Based on the two reconstruction errors of nodes, we define anomaly score function $f(\cdot)$ for node $v_i^{k'}$ to estimate its degree of abnormality as

$$s_i^{k'} = f(v_i^{k'}) = (1-\beta)\,\|\widehat{\mathbf{a}}_i^{k'} - \mathbf{a}_i^{k'}\|_2^2 + \beta\|\widehat{\mathbf{x}}_i^{k'} - \mathbf{x}_i^{k'}\|_2^2. \tag{2.10}$$

Based on the computed anomaly scores of nodes in the graph, the abnormality of the corresponding regions can be ranked in the image.

## 2.5 Experiments

In this Chapter, we first introduce three real-world datasets collected by our autonomous robot. Then we conduct experiments to evaluate the performance of SSGAE compared with several baseline methods. The experimental results are illustrated, including a comparison of performance, a parameter study, and an investigation into the effectiveness of its components.

### 2.5.1 Datasets

We evaluate SSGAE on three real-world datasets: LabPatrolling, BehaviorMonitoring, and AnoVisualGenome. The first two datasets, i.e., LabPatrolling and BehaviorMonitoring, are constructed from the human monitoring video clips collected by our autonomous robot in a real laboratory environment, which have been adopted in our previous work [7, 14, 21]. AnoVisualGenome is constructed by randomly selecting a subset of human-related images, which includes human activities in various environments, from a large-scale region caption dataset Visual Genome[1]. These three datasets consist of diverse region anomalies, i.e., single and contextual anomalies, and thus pose a challenge to detection algorithms. The instructions for these datasets are given as follows.

- **LabPatrolling** is constructed from the video clips when the mobile robot patrols around the laboratory. It includes various single anomalies, such as a man holding a baseball bat and a man holding an umbrella in the room, as well as a small number of contextual anomalies, such as a man making a phone call in the working area. It contains 5146 normal images for training, as well as 373 normal images and 21 abnormal images for testing.

- **BehaviorMonitoring** is constructed from another large-scale human monitoring dataset of video clips (approximately 100 hours). In this dataset, the mobile robot is navigated to several designated locations by a predefined program to monitor a variety of human activities taking place in the laboratory environment. It includes a wide range of contextual anomalies associated with many human behaviors. For instance, the behaviors of a man eating or sleeping in the working and resting areas are defined as normal and abnormal behaviors, respectively. It

[1] https://visualgenome.org/

contains 5548 normal images for training, as well as 585 normal images and 106 abnormal images for testing.

- **AnoVisualGenome** is constructed from Visual Genome [68] which provides dense annotations for regions on over 108K images. It includes several kinds of human activities in inappropriate environments as contextual anomalies, such as watching TV on the street and sitting on a couch on the beach. It contains 1427 normal images for training, as well as 218 normal images and 31 abnormal images for testing.

For our target task, after obtaining salient regions from images, we annotate region-level anomalies in the images, including anomalous human behaviors or irregular human-object interactions. Several normal and abnormal regions in the images are shown in Figure 2.1. We present further examples of normal and abnormal regions in the three datasets in Figure 2.5.

### 2.5.2 Experimental Setup

#### 2.5.2.1 Preprocessing

In the preprocessing stage, by utilizing advanced pre-trained deep models, we obtain regions with their captions in images and generate the visual and semantic features of regions to construct graphs.

Specifically, we utilize a dense captioning model Densecap[1] [1] pre-trained on Visual Genome [68] in a standard implementation to generate region candidates for the first two datasets and select the top-$n$ region candidates per image based on their confidence scores. By investigating the qualities of the generated regions with captions, $n$

[1] https://github.com/jcjohnson/densecap

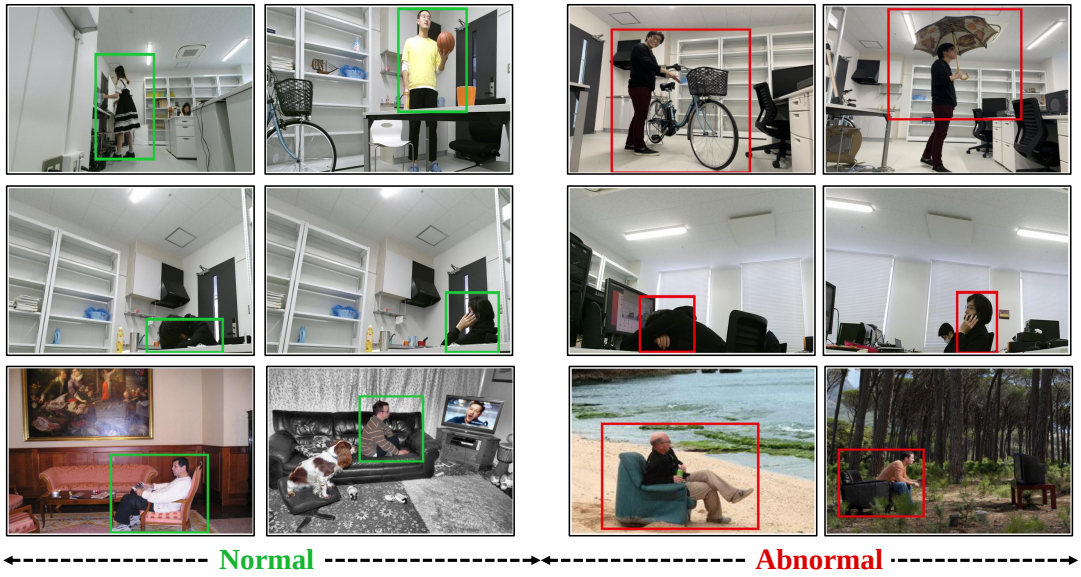



Figure 2.5: Examples of images with normal and abnormal regions. The normal and abnormal regions are shown in green and red boxes, respectively. The abnormal regions in the upper row are examples of single anomalies in LabPatrolling. In contrast to the normal regions in the middle and bottom rows, the abnormal regions in the same rows are examples of contextual anomalies in BehaviorMonitoring and AnoVisualGenome, respectively.

is set to 10 [7, 21, 40]. For AnoVisualGenome, as the number of ground-truth regions with captions per image ranges from 10 to 60, we randomly select 10 regions for each image.

Subsequently, ResNet101[1] [62] is adopted to extract the visual feature of each region from the output in the penultimate layer with dimension 2048. An SBERT [63] model named "all-mpnet-base-v2"[2] is adopted for transforming each region caption into an embedded vector with dimension 768. ResNet101 and SBERT are applied under their default settings and pre-trained on ImageNet [69] and 14 sentence datasets [63], respectively.

[1] https://pytorch.org/vision/stable/models/resnet.html
[2] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

#### 2.5.2.2 Baseline Algorithms

We evaluate the performance of SSGAE and compare it with several traditional and popular anomaly detection algorithms. These baseline methods include a traditional reconstruction-based algorithm, Auto-Encoders (AE) [70], a popular generative anomaly detection method, GANomaly[1] [61], two clustering-based region anomaly detection methods, Anomalous Image Region Detection (AIRD) [21] and Fast-and-Slow-Thinking Anomaly Detection (FSTAD) [7], as well as three variants of graph auto-encoders, Variational Graph Auto-Encoders[2] (VGAE) [66], Deep Anomaly Detection on Attributed Networks[3] (DOMINANT) [49], and Graph Attention Auto-Encoders (GATE) [59]. The instructions for the baseline algorithms are given as follows.

- **AE** [70] is a classical reconstruction-based method for anomaly detection. Both the encoder and the decoder are designed with fully-connected layers.
- **GANomaly** [61] is a popular generative anomaly detection method. It adopts an encoder-decoder-encoder module as a generator and three loss functions to jointly reconstruct images and features in a latent space.
- **AIRD** [21] is a one-class region anomaly detection method. It combines the visual, caption, and coordinate features of each region as its representation and employs an incremental clustering method to model normal regions.
- **FSTAD** [7] employs AIRD as its fast module for detecting single anomalies and devises a slow module recording neighboring regions with their visual features for detecting anomalous region pairs.

[1] https://github.com/samet-akcay/ganomaly
[2] https://github.com/DaehanKim/vgae_pytorch
[3] https://github.com/kaize0409/GCN_AnomalyDetection_pytorch

- **VGAE** [48] is the first model to extend the auto-encoder framework on graph data. It encodes node representations by GCN layers and utilizes an inner product decoder for reconstructing the adjacency matrix of graph data.

- **DOMINANT** [49] is the state-of-the-art graph auto-encoder for detecting anomalous nodes in attributed graphs by devising GCN-based components and adopting reconstruction errors as the anomaly scores.

- **GATE** [59] is a graph auto-encoder variant which stacks graph attention layers in its encoder and decoder for graph classification tasks.

#### 2.5.2.3 Implementation Details

We implemented SSGAE with the Pytorch[1] framework (version 1.6.0) on Ubuntu 18.04 equipped with a GPU of NVIDIA TITAN RTX (24 GB memory) and a CPU of i9-9820X. The objective function $\mathcal{L}$ is optimized by Adam [71] with a learning rate 0.004 and a weight decay $8 \times 10^{-5}$.

In the Spatial and Semantic Attributed Graph, the semantic similarity threshold $\theta_{\text{sim}}$ for building semantic edges is set to 0.5 in our experiments. In SSGAE, the attributed graph encoder is equipped with $L = 2$ hidden layers along with their MLP modules $\text{MLP}_{\text{Enc}}^{(l)}$, both of which contain two fully-connected layers with the hidden units $(2816 - 256 - 256)$ and $(256 - 256 - 128)$, respectively, with ReLU activation function. Accordingly, the graph attribute decoder also contains $L = 2$ hidden layers with their MLP modules $\text{MLP}_{\text{Att}-\text{Dec}}^{(l)}$, in which the fully-connected layers are symmetric to the layers in the encoder in terms of the number of their hidden units for reconstruction. In the graph structure decoder, the dimensions of the fully-connected layers in $\text{MLP}_{\text{Str}-\text{Dec}}$ are set to $(128 - 256 - 256)$. The hidden layers of other graph

[1] https://pytorch.org/

auto-encoder variants in the baselines are set to the same dimensions as SSGAE for a fair comparison. SSGAE and the other graph auto-encoder variants are trained for $T = 400$ epochs on the first two datasets and $T = 200$ epochs on AnoVisuaGenome. Hyper-parameter $\beta$ in SSGAE is set to 0.8, 0.8, and 0.9 for LabPatrolling, Behavior-Monitoring, and AnoVisuaGenome, respectively. When implementing other baseline methods, we retain the suggested settings in their original papers.

### 2.5.3 Experimental Results and Analysis

Figure 2.6 and Table 2.3 show the ROC curve and AUC score of SSGAE compared with the baselines on the three datasets, respectively. Moreover, Figure 2.7 illustrates the anomaly score distributions of all methods by boxplot, which displays the lower quartile, the median, and the upper quartile of the scores in a box and extends the box from the lowest to the highest scores by a line segment. We have the following findings based on the results.

1. SSGAE outperforms all the baseline methods on the three datasets and achieves $0.016-0.387$, $0.038-0.315$, and $0.043-0.345$ improvements in terms of their AUC scores on LabPatrolling, BehaviorMonitoring, and AnoVisualGenome, respectively. This validates the superiority of our method for the region anomaly detection task. The main reason is that SSGAE is capable of discriminating node representations from the Spatial and Semantic Attributed Graph and thus generates separated reconstruction errors to measure the abnormalities of regions, as shown in the example in Figure 2.8.

2. The previous methods, which do not consider region contexts, i.e., AE, GANomaly, and AIRD, achieve competitive performance on LabPatrolling, where most of the anomalies are single anomalies. This observation proves their effectiveness

in detecting single anomalies which are dissimilar to normal regions, e.g., normal and abnormal regions in the upper row in Figure 2.8. However, these methods do not perform well on BehaviorMonitoring and AnoVisual Genome, where there exist a large number of contextual anomalies. For instance, GANomaly achieves an AUC score of 0.911 on LabParolling, while it only achieves 0.794 and 0.687 on the other two datasets. The distributions of the anomaly scores on the two datasets shown in Figures 2.7(b) and 2.7(c) demonstrate that AE, GANomaly, and AIRD are unable to discriminate the normal and abnormal regions very well. We think the reason would be that without considering the region contexts, the contextual anomalies include similar human behaviors as normal regions, which are difficult to detect with these methods. To confirm the reason, we investigate the anomaly scores of the examples, including a normal region and a contextual anomaly, i.e., the no. 0 regions in the upper and bottom images in the left part of Figure 2.8. Compared with SSGAE, which outputs the anomaly score of $0.565/0.814$ on the normal/abnormal regions in Figure 2.8, AE, GANomaly, and AIRD output $0.425/0.462$, $0.199/0.381$, and $0.542/0.639$, respectively. These findings indicate that the methods which do not consider region contexts have deficiencies in detecting contextual anomalies compared with SSGAE.

3. Compared with other graph auto-encoder variants, SSGAE achieves significant performance gains with improvements of 0.043, 0.055, and 0.043 on the three datasets in terms of AUC scores. Accordingly, the anomaly scores of normal and abnormal regions generated by SSGAE are better discriminated compared with these baseline methods, as shown in Figure 2.7. The main difference between SSGAE and other graph auto-encoders is the sum aggregation strategy, which plays a critical role in discriminating the representations of node neighbors. We

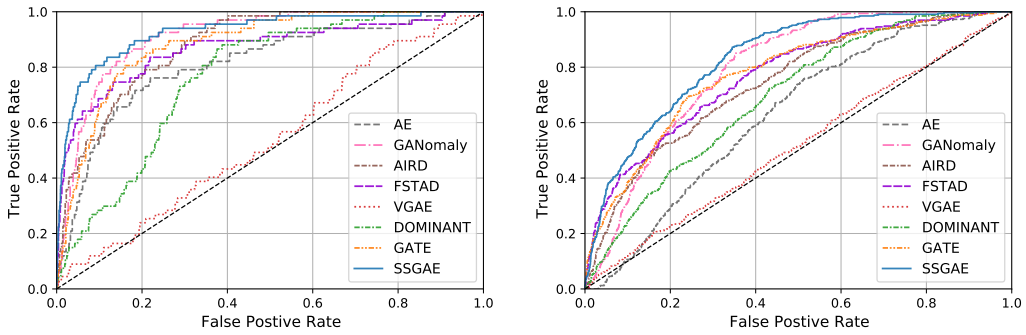
Table 2.3: AUC scores of SSGAE compared with the baseline methods.

| | Dataset | | |
|---|---|---|---|
| Method | LabPatrolling | BehaviorMonitoring | AnoVisualGenome |
| AE | 0.813 | 0.631 | 0.709 |
| GANomaly | 0.911 | 0.794 | 0.687 |
| AIRD | 0.881 | 0.745 | 0.794 |
| FSTAD | 0.868 | 0.772 | 0.701 |
| VGE | 0.540 | 0.517 | 0.524 |
| DOMINANT | 0.767 | 0.695 | 0.709 |
| GATE | 0.884 | 0.777 | 0.826 |
| **SSGAE**[1] | **0.927** | **0.832** | **0.869** |

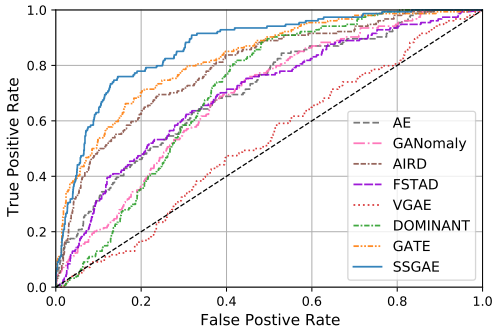
[1] The best performance of the method with AUC scores on the three datasets is in bold.

verify the effectiveness of the sum aggregation strategy in SSGAE by substituting it with the aggregation strategies in other graph auto-encoders, as illustrated in Chapter 2.5.5.

4. We observe that VGAE performs worst on the target task, although its encoder is similar to the encoders in other graph auto-encoders. We notice that compared with DOMINANT, GATE, and SSGAE, the decoder in VGAE only aims at reconstructing the graph structure without considering the reconstruction of node attributes in the graph. This fact implies that both the structure and the attribute reconstructions are necessary for our method of the task.



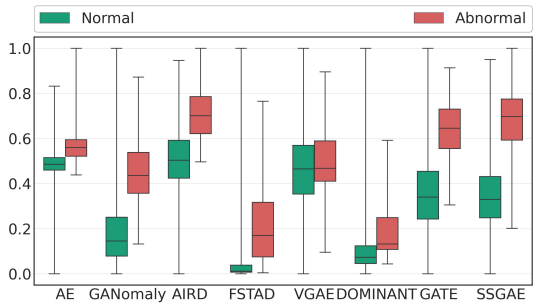

(a) LabPatrolling. (b) BehaviorMonitoring.



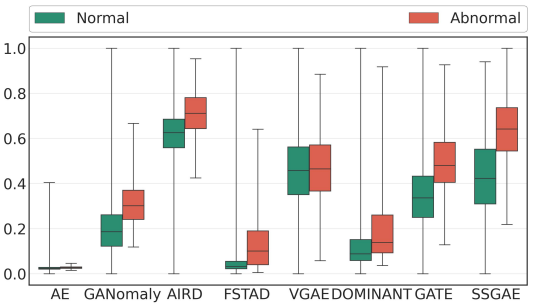

(c) AnoVisualGenome.

Figure 2.6: ROC curves of all methods on the three benchmark datasets.

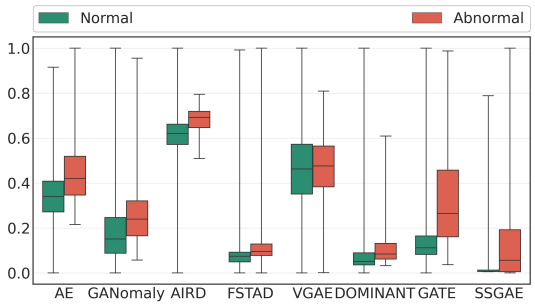
We also show an example of detecting normal and anomalous regions by SSGAE in Figure 2.8. In the upper image, the no. 0 region of a man making a phone call (the green box) in a resting area is normal, while the no. 0 region of the same behavior (the red box) in a working area in the bottom image is abnormal due to their different contexts. We visualize the original features of two regions and their embeddings generated by SSGAE with Principal Component Analysis (PCA) [72]. We see that although the two regions are closely located in the original feature space, trained on normal data, SSGAE can compress the two regions with their contextual information into well-separated embeddings and thus generate accurate anomaly scores in the right




(a) LabPatrolling.




(b) BehaviorMonitoring.




(c) AnoVisualGenome.

Figure 2.7: Distributions of anomaly scores on the three datasets.

part of Figure 2.8.

Considering the feasibility of applying our method to real-time region anomaly detection in human monitoring, we also evaluate the actual running time of the method in the test phase. For each test image, the proposed method outputs the anomaly scores of all regions with an average running time of 0.53s. We believe this performance is sufficient as we target human monitoring. Here we assume that the preprocessing procedure, which includes extracting pre-trained features and constructing graphs, is conducted before the monitoring process. The computation time of the preprocessing procedure during testing is about 3m48s, 7m58s, and 2m16s on LabPatrolling, BehaviorMonitoring, and AnoVisualGenome, respectively.

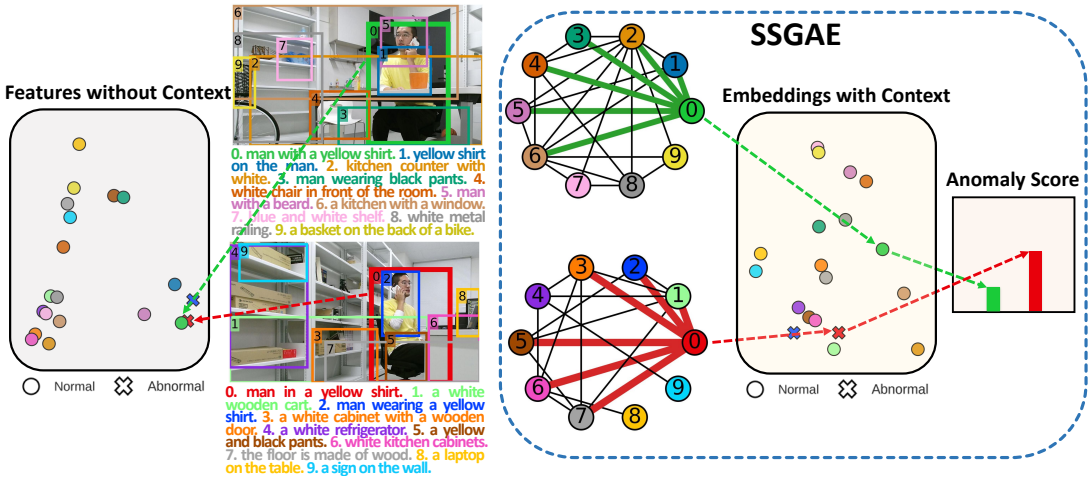


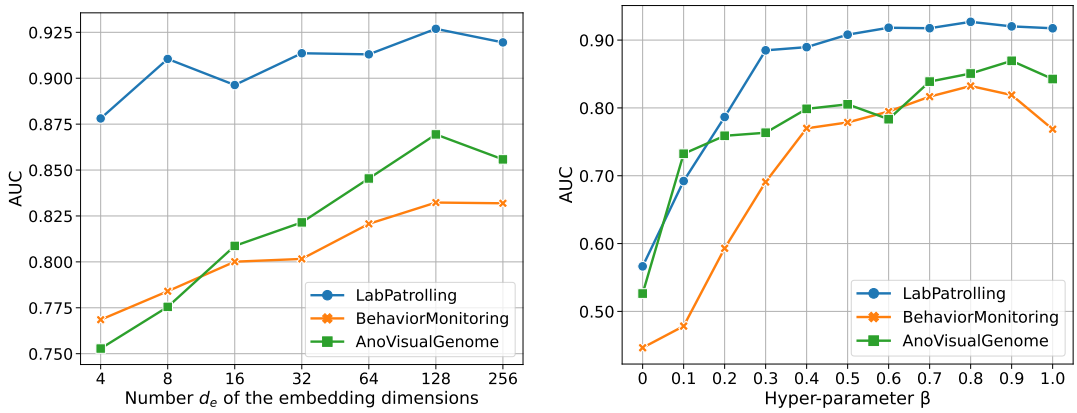

Figure 2.8: Example of detecting anomalous regions by SSGAE.

### 2.5.4 Parameter Sensitivity Study

To investigate the effects of embedding dimensions $d_e$ of the final hidden embedding and hyper-parameter $\beta$ in the objective function on the performance of SSGAE, we conduct experiments by modifying their values.

We first explore the sensitivity to dimension $d_e$ of the final hidden embedding by setting the values of $d_e$ from 4 to 256. We show the performance of SSGAE in Figure 2.9(a). On BehavoringMonitoring and LabPatrolling, the performance steadily improves when $d_e$ increases from 4 and reaches the peak value of 128, and then drops slightly when $d_e$ is 256. On AnoVisualGenome, the AUC score also steadily increases from $d_e = 4$ to $d_e = 128$. Then the performance gain becomes smaller when $d_e = 256$. These results show that $d_e$ should be in an appropriate range, e.g., from 64 to 256, for the target task.

We then modify the value of $\beta$ in the range of $\{0.0, 0.1, 0.2, \ldots, 1.0\}$ and show the results in Figure 2.9(b). According to the results, the AUC score rises when $\beta$ increases and reaches the peak value at 0.8, 0.8, and 0.9 on LabPatrolling, BehaviorMonitoring, and AnoVisualGenome, respectively. In particular, we can evaluate the performance




(a) Number $d_e$ of the embedding dimensions versus AUC.

(b) Hyper-parameter β in the objective function versus AUC.

Figure 2.9: Parameter sensitivity study of SSGAE.

of SSGAE only equipped with the structure decoder when $\beta = 0.0$ and only equipped with the attribute decoder when $\beta = 1.0$. We observe that our ablated model achieves poor results as it merely considers the structure reconstruction error, which indicates that attribute information is necessary for our task. On the contrary, by merely utilizing an attribute decoder in SSGAE, we cannot achieve the best results, which indicates the significance of jointly optimizing SSGAE by the structure reconstruction error and the attribute reconstruction error. These results show that it is necessary to find a trade-off to balance the two kinds of reconstruction errors for our task.

### 2.5.5 Effectiveness of Components

We further investigate the effectiveness of components in our method, i.e., the impacts of jointly considering the spatial and semantic relations in the proposed graph and the sum aggregation strategy in SSGAE.

We first conduct an ablation study by building two variants of the graph, i.e., the

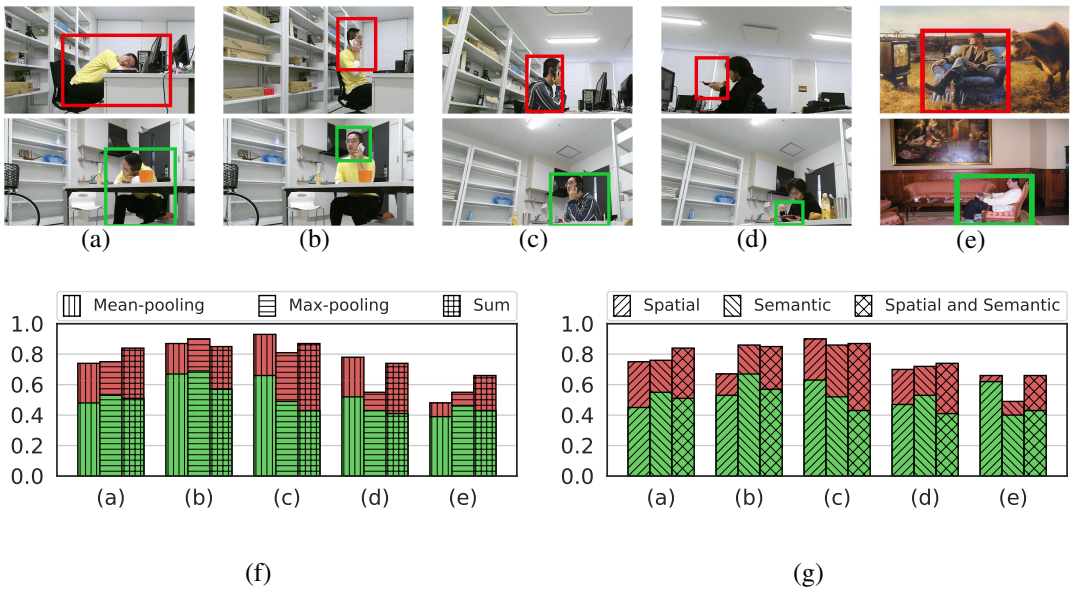
Table 2.4: The effectiveness of different components in our method.

| | Dataset | | |
|---|---|---|---|
| | LabPatrolling | BehaviorMonitoring | AnoVisualGenome |
| Spatial Attributed Graph | 0.915 | 0.807 | 0.833 |
| Semantic Attributed Graph | 0.924 | 0.778 | 0.791 |
| Mean-pooling Aggregation | 0.922 | 0.798 | 0.821 |
| Max-pooling Aggregation | 0.923 | 0.805 | 0.836 |
| **SSGAE**[1] | **0.927** | **0.832** | **0.869** |

[1] The best performance of the method with AUC scores on the three datasets is in bold.

spatial attributed graph and the semantic attributed graph which consider spatial relations only and semantic relations only among regions, respectively. Table 2.4 shows the results of SSGAE with these graphs. We observe that SSGAE on the spatial or semantic attributed graph achieves suboptimal performance, which implies the superiority of building both the spatial and semantic edges in the graph. We present several normal (green color) and abnormal (red color) examples in Figures 2.10(a)-2.10(e) with their anomaly scores in Figure 2.10(f). These examples in Figures 2.10(a)-2.10(e) include several human behaviors, such as a human sleeping, making a call, eating, and sitting on a couch, in different contexts. We observe that with the spatial attributed graph and the semantic attributed graph, the anomaly scores in Figure 2.10(f) of the normal and abnormal regions are not well-separated compared to SSGAE with the Spatial and Semantic Attributed Graphs. These results validate the effectiveness of the Spatial and Semantic Attributed Graphs on the target task.

We then verify the effectiveness of the sum aggregation strategy by substituting it




(f) (g)

Figure 2.10: Examples of abnormal and normal regions with anomaly scores. (a)-(d): examples of abnormal regions with red boxes and normal regions with green boxes in a laboratory environment. (e): examples of an abnormal region with a red box outside a room and a normal region with a green box inside a room. (f): anomaly scores of the abnormal regions with red color and normal regions with green color in (a)-(e) by the different kinds of graphs. (g): anomaly scores of the abnormal regions with red color and normal regions with green color in (a)-(e) by the different aggregation strategies.

with the mean-pooling and the max-pooling strategies in SSGAE. Based on the results in Table 2.4, SSGAE adopting the mean-pooling or max-pooling aggregation strategy achieves competitive performance on LabPatrolling. The reason would be that most anomalous regions in LabPatrolling are single anomalies and are thus easy to be detected by any of the aggregation strategies. However, the diverse contextual anomalies in BehaviorMonitoring and AnovisualGenome need to be judged by combining the regions with their contexts. Figure 2.10 shows the anomaly scores of regions in Figures 2.10(a)-2.10(e) with different strategies in Figure 2.10(g). We observe that SSGAE adopting the sum aggregation strategy discriminates the normal and abnormal regions better than SSGAE adopting the other two strategies in terms of their anomaly

scores. For instance, the normal and abnormal regions in Figure 2.10(a) show a human sleeping in the working and resting areas. SSGAE with the sum aggregation strategy generates the highest anomaly score for the abnormal region and a relatively low score for the normal region in Figure 2.10(a) compared to SSGAE with the other two strategies. This implies the effectiveness of adopting the sum aggregation strategies in SSGAE for detecting contextual anomalies in our task.

## 2.6 Summary

In this Chapter, we tackled the region anomaly detection task in human monitoring by constructing the Spatial and Semantic Attributed Graph and devising the graph auto-encoder framework SSGAE. To characterize the anomalous region based on its content and context, we built the graph to model regions with their spatial and semantic relations in the image. Subsequently, SSGAE which is equipped with the sum aggregation strategy [26] and consists of one encoder and dual decoders, was introduced for our task. Due to the lack of rare and diverse anomalies in human monitoring, SSGAE is trained to reconstruct the node attributes and structures in the graph in a one-class anomaly detection manner. In the test stage, the structure and the attribute reconstruction errors are then jointly employed in the anomaly score to estimate the abnormality of nodes as well as their corresponding regions. We conducted extensive experiments and analyzed the results to evaluate the superiority of SSGAE on the target problem.

# Chapter 3

# Phrase Similarity Graph for Judging Credible and Unethical Statistical Data Explanations

## 3.1 Overview

As the impact and the presence of AI systems on our societies increase, their unethical misconducts are prone to severe reproach. The misconducts of Deepfakes pose a serious threat to truth, trust, and privacy by spreading false information and manipulating public opinions [73]. The hijacking event of the chatbot Tay clearly shows that pure benevolence could turn into an opposite outcome [74], e.g., inflammatory tweets by a chatbot are often unethical, and harm the reputation of its producer. Moreover, although the advent of ChatGPT[1] has the potential to revolutionize various industries and aspects of our daily lives [75], such a practical large language model also holds the possibility of generating and spreading seemingly convincing yet biased informa-

[1] https://openai.com/blog/chatgpt

tion [76,77], such as fake news. These kinds of information pose a significant challenge to the morality of our society. Among such misinformation, those that are credible are more influential than others, as their contents are more likely to be believed by people.

In this Chapter, among such reasons, we tackle exploitation of human instincts in statistical data explanation. Rosling et al.'s book "Factfulness" has known a global success and emphasizes the importance of thinking based on facts and correct understandings [2]. The book includes examples of unethical and biased explanations each of which is denied by the accompanied statistical data. We, however, argue that such a thinking attitude is not always adopted and even accepted. Take as an example an explanation "Asia is the cause of the large amount of CO2 emissions"[1] with its statistical data depicting GDP per capita, total amount of CO2 emissions, and CO2 emissions per capita of four continents in Figure 3.1. The statistical data show that although Asia seems to be the cause in the view of total emissions in the first plot, the explanation is refuted by the per-person emission view with respect to the GDP per capita in the second plot. However, due to the single perspective instinct, i.e., our tendency to prefer a single cause or solution [2], some portion of people would believe the explanation, even though the statistical data clearly contradicts it. Such an unethical explanation deserves special attention as it highlights challenges to our rationality and understanding. In this Chapter, we are going to define 21 types of such credible and unethical explanations each with its statistical data.

Moreover, we provide countermeasures to such explanations. In Chapter 3.5, we first devise three methods $\alpha$, $\beta$, and $\gamma$ for judging whether an explanation is credible and unethical based on phrase embedding and carefully designed conditions. The phrase embedding technique is an extension of word embedding [64, 78–81] to project

[1] All unethical examples in this thesis are either adopted from other sources or slightly modified from them and do not reflect the beliefs of the author nor our organizations. In all cases, such examples are not believed by the authors of the sources, either.

phrases in a high dimensional vector space, where semantically similar phrases are embedded near each other. The conditions in the three judgment methods are designed to compare semantic relevance between phrases specified in the explanations. We conduct experiments on the statistical data explanations to evaluate the effectiveness of the three methods.

Based on the experimental results, only method $\beta$ achieves relatively low accuracy on the target task due to numerous counter-intuitive semantic similarities between phrases in the designed conditions. To address the limitation and improve the accuracy of method $\beta$, in Chapter 3.6, we propose a new graph-based method $\beta^2$. Method $\beta^2$ first constructs a Phrase Similarity Graph to model the statistical data explanation by considering more phrases. The graph can explicitly represent these phrases and their semantic similarities, where the conditions for the judgment can be simply generated based on node combinations. Then a credibility score for judging the credibility of the explanation is proposed based on the generated conditions and graph entropy.

The main contributions of this Chapter as summarized as follows.

- We define 21 types of credible and unethical explanations with the exploitation of Rosling et al.'s ten human instincts [2], each of which is accompanied by its statistical data.

- We devise three methods $\alpha$, $\beta$, and $\gamma$ for judging credible and unethical statistical data explanations. The three judgment methods investigate the credibilities of the unethical explanations by comparing semantic relevance degrees between specified phrases in the explanations.

- To address the limitation of method $\beta$, we propose a graph-based judgment method $\beta^2$. Method $\beta^2$ constructs a Phrase Similarity Graph to consider more

phrases for generating necessary conditions and adopts graph entropy to quantify the different importance of the generated conditions for more accurate judgment.

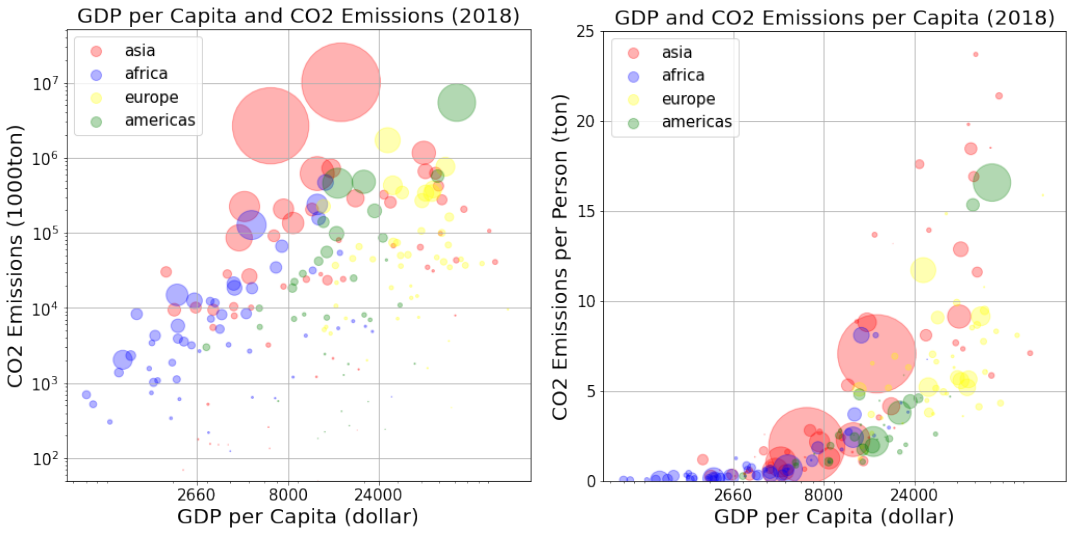



Figure 3.1: Statistical data of GDP per capita and the total amount of CO2 emissions versus GDP per capita and CO2 emissions per capita.

## 3.2 Related Work

### 3.2.1 Unethical Explanations

Unethical and biased explanations are widely generated in diverse fields around the world [82], such as fake news and misinformation. Misinformation can be defined as incorrect or counterfactual information, while fake news is a specific type of misinformation which is intentionally created to mislead the audience [82]. Detecting fake news is a challenging Natural Language Processing (NLP) task involving two problems: characterization and detection [83]. Considering feature selection and extraction, Reis et al. [84] designed informative features, which consider semantic and syntactic

properties, political biases, credibility, and environments of news, for automatic detection of fake news. Vlachos et al. [85] introduced fact-checking tasks and discussed baseline approaches to assess truthfulness of explanations by measuring their semantic similarities. Detecting fake news is usually formulated as a classification task in a supervised manner [86, 87]. Through integrating meta data with texts, a hybrid Convolutional Neural Network (CNN) is devised to classify fake news based on surface-level linguistic patterns [88]. Moreover, since fake news with images or videos is becoming increasingly prevalent with the development of multimedia technology, multimodal information including visual and textual features has been explored for more accurate detection [89–91].

In this Chapter, we limit our attention to explanations of statistical data and focus on their unethical nature and credibility due to instinct exploitation. Statistical ethics refers to the ethical consideration and principles which guide the collection, analysis, interpretation, and communication of statistical information [92]. Statistical ethics covers a wide range of topics, such as the selection bias in data collection for clinical research [93], the misuse and abuse of statistical data for biomedical research [94], and the survivorship bias in statistical for longitudinal mental health surveys during the COVID-19 pandemic [95]. These works mainly focus on addressing ethical concerns in statistical data, aiming to promote the integrity and responsible use of data in their domains. Different from these works, we argue that credible and unethical explanations of statistical data due to human instinct exploitation deserve special attention since they can lead to the formation of stereotypes and prejudice against people. Such explanations may hinder people from developing correct understandings of the facts even if statistical data support them. To the best of our knowledge, no previous work tackles the problem of judging credible and unethical explanations on statistical data with AI methods. Our work is the first one to define and investigate such explanations

through AI techniques.

### 3.2.2 Semantic Similarity-Based Methods for NLP Tasks

As we explained in Chapter 3.1, our methods $\alpha$, $\beta$, and $\gamma$ are based on phrase embedding and carefully designed comparisons to judge the credibility of the statistical data explanation. Measuring semantic similarity between various text components such as words, sentences, or documents has been explored in a wide range of downstream NLP tasks, such as machine translation [96], information retrieval [97] and question answering [98]. Li et al. [99] measured semantic similarity between words using multiple information sources, including attributes path lengths, depths, and local densities in a hierarchical semantic knowledge base. To reduce the ambiguity in words, a robust semantic similarity measure [100] was devised by utilizing information including page counts and lexico-syntactic patterns from text snippets of a Web search engine. Similar to [100], Normalized Google Distance (NGD) [101] was proposed to measure the similarity between two terms based on query results of Google search engine.

Semantic similarity methods have exploited the recent developments in neural networks and word embedding to enhance their performance [102]. In contrast to adopting traditional static word embedding [79, 103] for semantic similarity measurement between words [104], contextualized word embedding generated from modern neural language models, such as ELMo [80], GPT-2 [81], and BERT [64], has been widely employed for semantic similarity tasks [105]. The latter approach possesses over the former an advantage of capturing rich syntactic and semantic properties of words under diverse linguistic contexts. Moreover, for semantic similarity tasks between two sequences of multiple words, such as phrases and sentences, InferSent [106] employs a bi-directional Long-Short Term Memory (LSTM) with a max-pooling operator as a

sentence encoder to generate sentence embedding. Trained on a number of natural language prediction tasks, Universal Sentence Encoder [107] models the meaning of word sequences to encode sentences into high dimensional vectors. Sentence-BERT [63] adopts Siamese and triplet architectures based on the pre-trained BERT network to generate semantically meaningful embedding for sentences. Furthermore, the semantic similarities of sentences can be directly compared with cosine-similarity between their embeddings.

### 3.2.3 Graph-Based Methods for Misinformation Detection

To address the limitation of method $\beta$, we propose a graph-based method $\beta^2$. Graph structures have been widely employed in fact-checking and misinformation detection, as they can make the structure of free text explicit and easily manageable by downstream algorithms. These works can be mainly classified into similarity-based and knowledge-based approaches. Similarity-based approaches often represent social media posts [108], sentences, or words in news articles [109–111] as nodes and build edges to represent their relations in a graph. TextRank [112] is adopted to identify credible statements from a graph in which the sentences and their semantic similarities represent nodes and edges [109], respectively. Utilizing the same kind of graph, Biased TextRank [110] associates an explanation extraction with a fact-checking task by comparing the similarities between the extracted statements with the ground truth. On the other hand, knowledge-based approaches often retrieve evidence which supports or refutes the information from a large and reliable knowledge graph [83, 113]. Vedula et al. [114] jointly exploit concept-relationship structures and semantic contextual cues from a knowledge graph to detect the veracity of an input fact and generate a human-comprehensible explanation justifying the fact. For health misinformation

detection, a knowledge-guided graph attention network is devised by incorporating a medical knowledge graph and an article-entity bipartite graph [115]. Different from these graph-based methods for misinformation detection tasks, the Phrase Similarity Graph in our method considers more phrases and their semantic similarities to address the issue of counter-intuitive semantic similarities, which improves the accuracy for judging the statistical data explanations.

We adopt sub-graph entropy in our graph-based method $\beta^2$. Graph entropy is a measure to understand and analyze the structure and complexity of a graph, which is often utilized to quantify the degree of uncertainty for graph data. Graph entropy is usually task-specific, i.e., it depends on the characteristics of the network. These works include structure and feature entropy for node embedding dimension selection [116], parametric graph entropy for analyzing information processing [117], and conditional substructure entropy for graph anomaly detection [118]. Among such works, Sen et al. [119] define the sub-graph entropy by focusing on the complexity of connections between nodes in functional brain networks. The sub-graph entropy is computed by exploring the node connectivity, i.e., edge weights, to evaluate the importance of each sub-graph in a whole graph. Since the Phrase Similarity Graph in method $\beta^2$ considers node combinations and their connections from its sub-graphs to generate comparison conditions for judgment, we utilize sub-graph entropy to measure the importance of the comparison conditions from different sub-graphs.

## 3.3 Problem Formulation

### 3.3.1 Rosling et al.'s Ten Human Instincts

As we stated in Chapter 3.1, we focus our attention on the credible and unethical explanations of statistical data with the exploitation of Rosling et al.'s ten human instincts [2]. The ten instincts are listed below, which could be considered as innate, typically fixed patterns of human thinking.

(1) The gap instinct: our tendency to divide all kinds of things into two distinct and often conflicting groups, with an imagined, huge gap in between.

(2) The negativity instinct: our tendency to notice the bad more than the good.

(3) The straight line instinct: our tendency to believe that the increase is a straight line.

(4) The fear instinct: our tendency to focus our attention on what we are afraid of.

(5) The size instinct: our tendency to misjudge the size of things or the importance of a single number/instance.

(6) The generalization instinct: our tendency to categorize and generalize things all the time.

(7) The destiny instinct: our tendency to consider that several things never change due to their innate characteristics.

(8) The single perspective instinct: our tendency to prefer a single cause or solution.

(9) The blame instinct: our tendency to find a clear, simple reason for why something bad has happened.

(10) The urgency instinct: our tendency to want to take an immediate action in the face of a perceived imminent danger.

### 3.3.2 Judging Credible and Unethical Statistical Data Explanations

We assume the following five conditions for the credible and unethical explanations of statistical data.

(1) Data seem to be valid, ideally taken from an authoritative source, e.g., WHO.

(2) The explanation is significant.

(3) The explanation seems to be believed by a certain number of people.

(4) The data can prove why the explanation is not valid.

(5) The explanation exploits at least one of the ten human instincts in Chapter 3.3.1.

Conditions (1), (4), and (5) contribute to the unethical nature of a statistical data explanation, which consider its validity, objectiveness, and exploitation of human instincts, respectively. As we are going to consider variants for each explanation by replacing its phrases while keeping the data, we have inserted the phrase "seem to" in condition (1). Condition (4) assures that we can refute the explanation based on the accompanied data only. Without this condition, the unethical nature of the explanation depends on the beliefs and the knowledge of the judge, which are diverse. The ten instincts mentioned in condition (5) are prone to unethical notions such as segregation, prejudice, fear, and inequality. Conditions (2) and (3) are also necessary as they consider the significance and the credibility, respectively. Condition (3) also needs the phrase "seem to" as it depends on subjectivity and knowledge of various people.

Without (2) and (3), the explanation is not harmful as people do not pay attention to them.

As we discussed in Chapter 3.1, unethical statistical data explanations which are credible deserve more attention than those that are not because they have a greater negative impact on correct human understanding. Therefore, our target problem is to judge whether a given explanation is credible and unethical (class 1) or not (class 0).

The target problem is formulated as the test phase of a binary classification task, where the goal is to predict the class labels of the explanations. The ground-truth class labels are given by humans for evaluation purposes only. The input of the target problem is an explanation, its statistical data, and its phrases, which will be explained in Chapter 3.5. The output is the predicted class label (0 or 1) of the explanation. To evaluate our judgment methods, we utilize accuracy as the evaluation metric.

## 3.4 21 Types of Statistical Data Explanations

We define 21 types (I-XXI) of credible and unethical explanations, which describe 7 kinds of statistical data. The data are (A) values of a probabilistic variable under 2 conditions, (B) a scatter plot of 2 probabilistic variables, (C) scatter or bar plots in different categories or times, (D) a probability density function of a probabilistic variable and a plot of its average value, (E) a time-series chart or scatter plots in chronological order, possibly with an additional one, (F) scatter plots of 2 probabilistic variables focusing on the total values and the average values, and (G) a funnel plot. Examples of the statistical data are shown in Figures 3.2, 3.3, and 3.4.

For each type of explanation, we code the exploited instincts and its statistical data. For example, in type I explanation, A-2 represents that the explanation exploits instinct (2) to explain the statistical data (A). $X$ and $Y$ are phrases which are respectively spec-

ified as a subject and its characteristics in the explanation. In addition, we clarify why the explanation is not valid according to its statistical data. Lastly, we provide candidates of phrases $X$ and $Y$ to generate variants of each type. The variants are generated by replacing phrases $X$ and $Y$ in the original explanation with the provided candidates.

(Type I) A-2, A-4: Deep-fried food boosts pancreatic cancer risk.
$X$: deep-fried food. $Y$: pancreatic cancer.

(Clarification) The relative risk of pancreatic cancer is only increased by 0.25% [3]. Statistically testing the difference between the two groups will fail.

(Candidates for variants) $X$: alcohol abuse, heavy drinking, long-distance running. $Y$: Alzheimer's disease, periodontal disease, flu, alopecia areata, bone fracture, nosebleeds.

We have two variations for type II explanation, which are used for upper-left and lower-right countries in Figure 3.2 II.
(Type II-1) B-8: Cuba is the poorest of the healthiest countries.
$X$: Cuba. $Y$: poorest.
(Type II-2) B-8: United Arab Emirates (UAE) is the richest of the unhealthiest countries.
$X$: United Arab Emirates. $Y$: richest.

(Clarification) (Type II-1) Cuba is also the healthiest of the poorest countries. It is inappropriate to consider only one side.
(Type II-2) The same reason applies to UAE, which is also the unhealthiest of the richest countries.

(Candidates for variants) (Type II-1) $X$: Bangladesh, North Korea, Nicaragua. $Y$: richest.

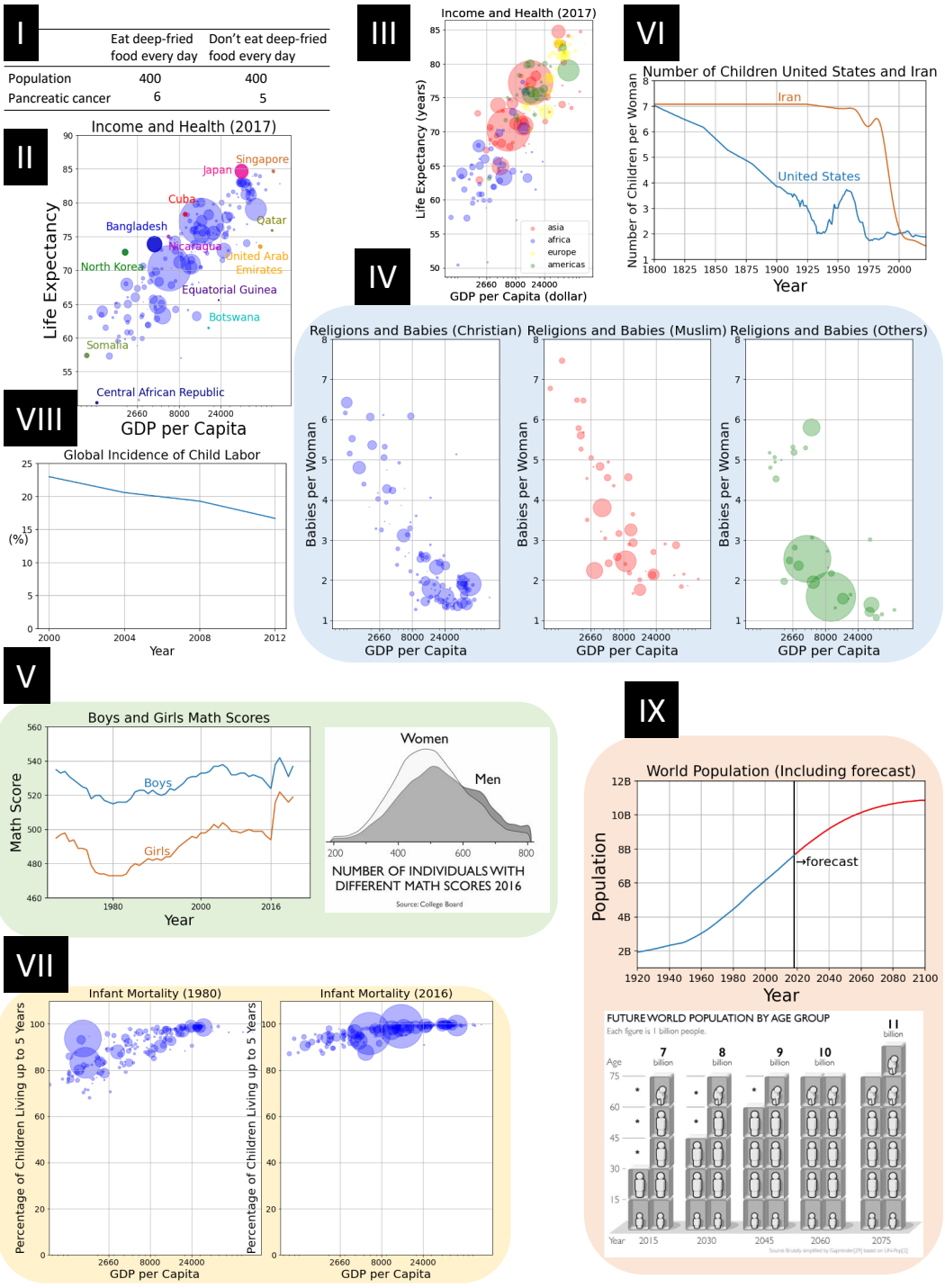



Figure 3.2: (Best in color) Statistical data in explanations (I-IX). Data are adopted or modified from [2], [3], or Gapminder [4].

(Type II-2) $X$: Qatar, Equatorial Guinea, Botswana. $Y$: poorest.
Note that Cuba and UAE are respectively compared among the healthiest and unhealthiest countries in these explanations.

(Type III) B-3: Life expectancy continues to grow in proportion to GDP per capita.
$X$: life expectancy. $Y$: proportional to GDP.

(Clarification) Note that the horizontal axis in Figure 3.2 III is set to a logarithmic scale, which is non-linear. The average life has an upper bound.

(Candidates for variants) $X$: healthy life expectancy. $Y$: inversely proportional to GDP, not correlated to GDP.

(Type IV) C-1, C-6, C-7, C-8: Muslims have many babies compared to Christians.
$X$: Muslims. $Y$: many babies.

(Clarification) All the 3 plots show that the number of babies decreases as the income increases, and there is no significant difference in the distribution. In fact, the average number of children per woman is 3.1 among Christians and 2.7 among Muslims.

(Candidates for variants) $X$: Judaisms, Christians. $Y$: few babies.

(Type V) D-1, D-2, D-6, D-8: Women have lower math scores than men.
$X$: women. $Y$: low math score.

(Clarification) The left plot shows that girls (women) have lower average scores than boys (men). However, the right plot shows that there exists an almost complete overlap between the two groups.

(Candidates for variants) $X$: men. $Y$: high math score, low English score, high English score.

(Type VI) E-7, E-8: Iranians have many children compared to Americans in the 21st century.

$X$: Iranians. $Y$: many children.

(Clarification) In the past centuries, Iranians had more children than Americans. In this century the two groups are similar in the number of children.

(Candidates for variants) $X$: Afghans, Americans, French. $Y$: few children.

(Type VII) E-1, E-6, E-7, E-8: Infant mortality rates in developing countries are still significantly higher than in advanced countries.

$X$: developing countries. $Y$: high infant mortality rates.

(Clarification) The percentage of children living up to 5 years is now over 85% in most countries, and there is no significant difference between advanced and developing countries.

(Candidates for variants) $X$: advanced countries. $Y$: low infant mortality rates, low enrollment rates, high enrollment rates.

(Type VIII) E-2, E-5: Child labor is about 15% and is not decreasing.

$X$: child labor. $Y$: not decreasing.

(Clarification) The percentage of child labor is decreasing.

(Candidates for variants) $X$: child hunger, child mortality. $Y$: increasing, decreasing, not increasing, constant.

(Type IX) E-3: The world's population will just increase.

$X$: world population. $Y$: will just increase.

(Clarification) The bottom plot shows that the populations of younger generations

are stable and those of older ones slowly increase. As the results, the population growth will be controlled.

(Candidates for variants) $Y$: will rapidly increase, will just decrease, will rapidly decrease, will keep constant.

(Type X) E-2, E-4: Since year 2000, compared to 1980, there is an increasing in natural disasters and an increasing in deaths from natural disasters.
$X$: increasing in natural disasters. $Y$: increasing in deaths from natural disasters.

(Clarification) The number of natural disasters is increasing, whereas the number of deaths from disasters is fluctuating and tends to decrease.

(Candidates for variants) From this type, we use $\{\}$ as there are many candidates. A variant should discuss one of the three topics. $X$: $\{$increasing in, decreasing in, constant$\}$ $\{$natural disasters, epidemic damages, industrial accidents$\}$. $Y$: $\{$increasing in, decreasing in, constant$\}$ deaths from $\{$natural disasters, epidemic damages, industrial accidents$\}$.

(Type XI) E-2, E-5, E-10: The death of many babies (4 million) is increasing.
$X$: death of many babies. $Y$: increasing.

(Clarification) Nearly 10 million babies died 40 years ago, but recently the number has fallen to 4 million and the situation is improving.

(Candidates for variants) $X$: death of many $\{$children, adults, old people$\}$. $Y$: not decreasing, decreasing, not increasing, constant.

(Type XII) F-1, F-6, F-7, F-8, F-9: Asia is the cause of the large amount of CO2 emissions.
$X$: Asia. $Y$: large amount of CO2 emissions.

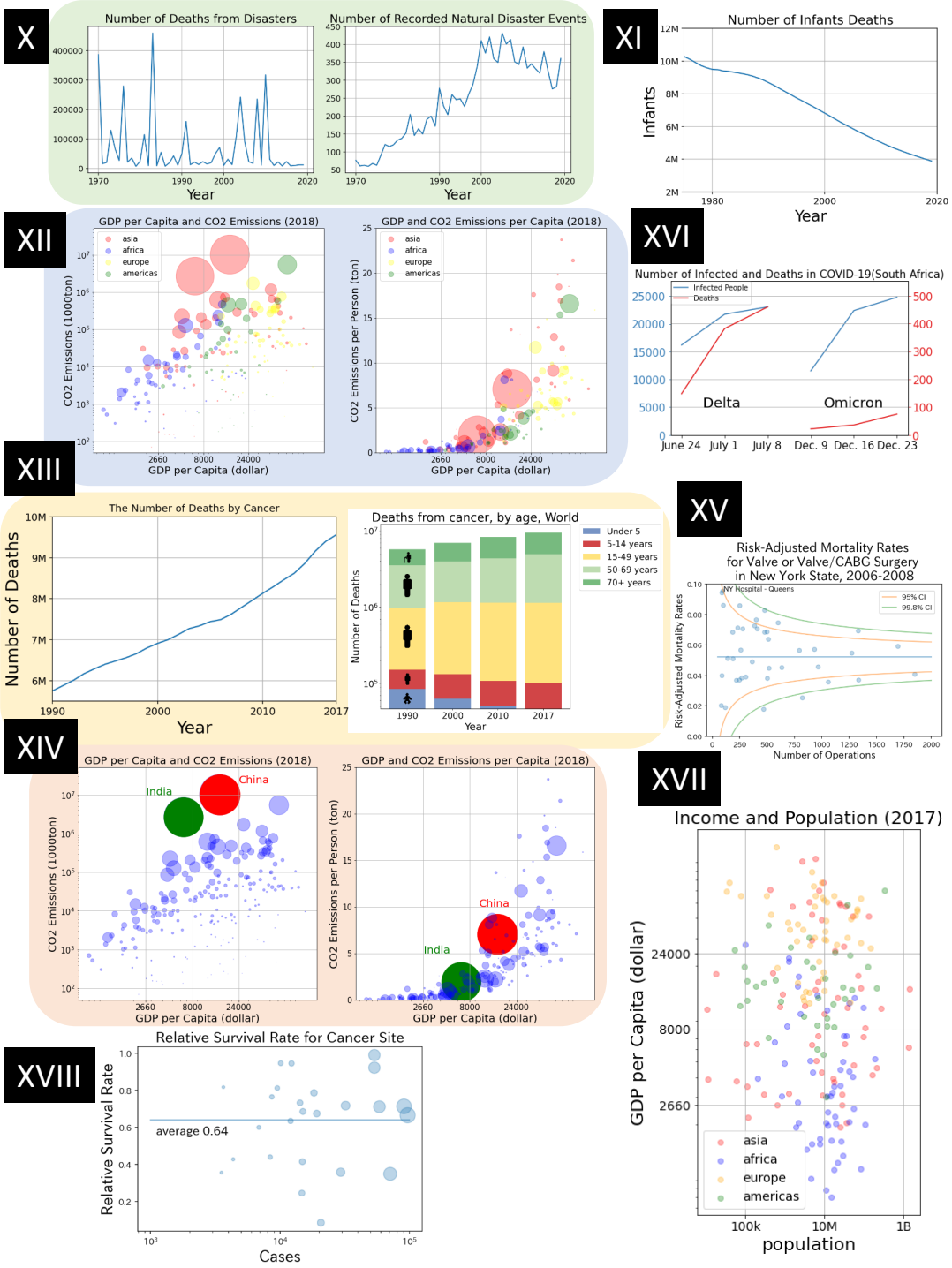



Figure 3.3: (Best in color) Statistical data in explanations (X-XVIII). Data are adopted or modified from [2], [3], or Gapminder [4].

(Clarification) Asian countries seem to be the cause in the view of total emissions, which is denied by the per person emission view with respect to the GDP per capita.

(Candidates for variants) $X$: Africa, Europe[1].
$Y$: small amount of CO2 emissions.

(Type XIII) E-2, E-8: The risk of death from cancer is increasing worldwide.
$X$: risk of death from cancer. $Y$: increasing.

(Clarification) The number of deaths from cancer is increasing, which is the result of the increase of the elderly in number.

(Candidates for variants) $X$: risk of death from $\{$Alzheimer's, heart$\}$ disease. $Y$: decreasing, constant.

(Type XIV) F-8, F-9, F-10: China is the cause of the large amount of CO2 emissions.
$X$: China. $Y$: large amount of CO2 emissions.

(Clarification) A large population inevitably leads to an increase in CO2 emissions. In terms of CO2 emissions per person, the explanation is denied.

(Candidates for variants) $X$: United Kingdom, India, United States. $Y$: small amount of CO2 emissions.
Note that this type gives a more precise view than type XII, which explains continents.

(Type XV) G-1, G-8, G-9: Small hospitals are dangerous hospitals[2]. $X$: small hospitals. $Y$: dangerous hospitals.

(Clarification) The funnel plot shows that most of the data points are within the confidence interval [3]. Thus there is no such tendency.

[1] We omitted Americas, which is diverse.
[2] We repeated the word "hospitals" to correctly measure the relevance between $X$ and $Y$.

(Candidates for variants) $X$: large hospitals. $Y$: safe hospitals.

(Type XVI) E-1, E-6, E-7, E-8: Omicron strain of COVID-19 is less dangerous than Delta strain.

$X$: Omicron strain. $Y$: less dangerous.

(Clarification) Judging the dangerous degree of Omicron strain only by the number of deaths is inadequate. Omicron strain is more dangerous than Delta strain in the view of infections.

(Candidates for variants) $X$: Alpha strain, Beta strain, Delta strain, Gamma strain. $Y$: more dangerous.

(Type XVII) B-1, B-6, B-7: Africa has lower GDP per capita than other regions.

$X$: Africa. $Y$: low GDP.

(Clarification) Not all African countries have lower GDP per capita than other regions.

(Candidates for variants) $X$: Asia, Americas, Europe. $Y$: high GDP.

(Type XVIII) B-2, B-6, B-7, B-8: The average 5-year survival rate for cancer is 64% so short life expectancy is predicted than other diseases.

$X$: cancer. $Y$: short life expectancy.

(Clarification) The explanation is an overgeneralization because the survival rates for several less dangerous cancers are higher.

(Candidates for variants) $X$: Alzheimer's disease, periodontal disease, heart disease, pneumonia. $Y$: long life expectancy.

We hesitated between "short" and "long" in $Y$ but finally chose the former as it is more credible than the latter. The term "than other diseases" has been added as without it

one would compare cancer patients with people with no disease. Note that this type is kept to show the difficulty of the target problem despite its flaw, i.e., the statistical data do not contain survival rates of other diseases, which reflects the difficulty of discussing the remaining diseases all at once.

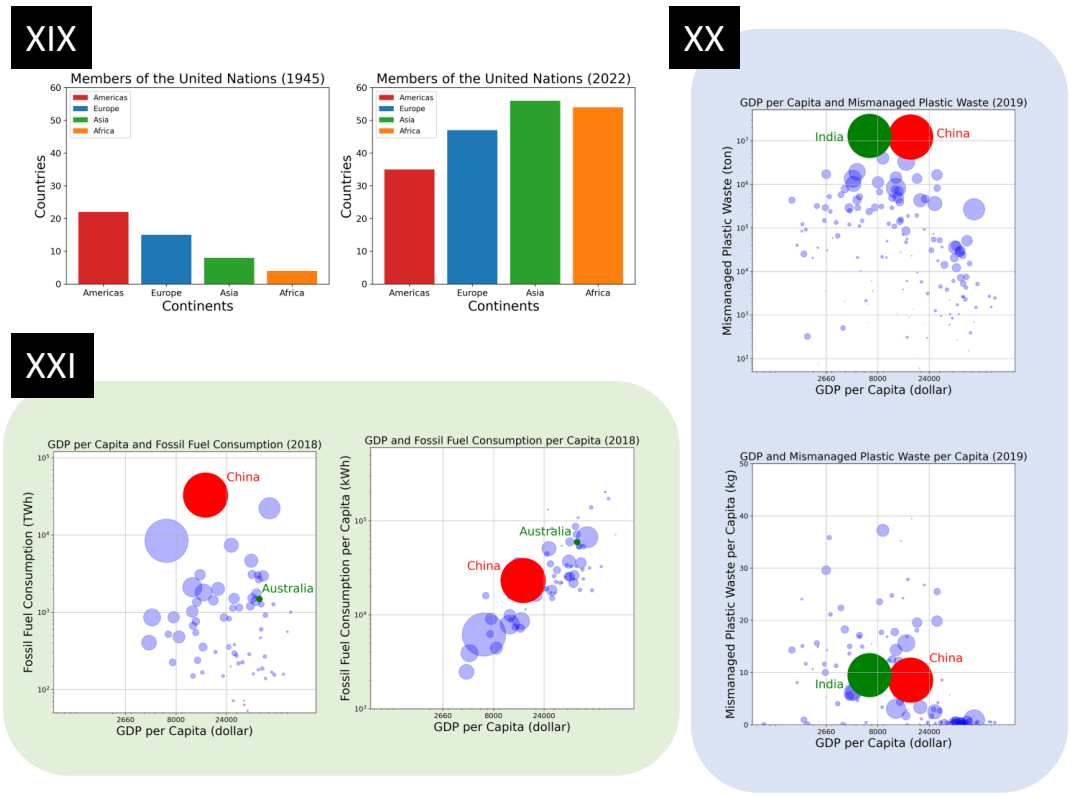



Figure 3.4: (Best in color) Statistical data in explanations (XIX-XXI). Data are adopted or modified from Gapminder [4], or Our World in Data[1].

(Type XIX) C-1, C-7, C-8: Americas have more members of the United Nations than Africa.

$X$: Americas. $Y$: many members of the United Nations.

(Clarification) In the last century, Americas and Europe had more members of the

[1]https://ourworldindata.org/

United Nations compared with Asia and Africa. Currently, the United Nations has a truly global coverage. The number of members of the United Nations in Asia or Africa is almost the same as the number in Americas and Europe.

(Candidates for variants) $X$: Europe, Asia, Africa. $Y$: few members of the United Nations.

(Type XX) F-8, F-9, F-10: India is the cause of the large amount of mismanaged plastic waste.

$X$: India. $Y$: large amount of mismanaged plastic waste.

(Clarification) Countries with large populations, such as China and India, inevitably lead to an increase in mismanaged plastic waste. In terms of mismanaged plastic waste per person, the explanation is denied.

(Candidates for variants) $X$: China, United Kingdom, United States. $Y$: small amount of mismanaged plastic waste, large amount of plastic emissions, small amount of plastic emissions.

(Type XXI) F-1, F-7, F-8: Australia has lower fossil fuel consumption than China.

$X$: Australia. $Y$: low fossil fuel consumption.

(Clarification) Australia seems to have lower fossil fuel consumption than China in the view of total consumption, which is denied by the per-person consumption view with respect to the GDP per capita.

(Candidates for variants) $X$: United Kingdom, United States. $Y$: high fossil fuel consumption.

## 3.5 Phrase Embedding-Based Judgment Methods

Based on the subjects and characteristics, i.e., phrases $X$ and $Y$, the 21 types of explanations can be classified into three categories including $(\alpha)$ habits and diseases, $(\beta)$ subjects and properties, and $(\gamma)$ subjects and trends. Following this criteria, type I belongs to $(\alpha)$ category. Types II, IV-VII, XII, and XIV-XXI belong to $(\beta)$ category. Types III, VIII-XI, and XIII belong to $(\gamma)$ category.

Accordingly, we devise three methods $\alpha$, $\beta$, and $\gamma$ to judge the explanations of their respective categories based on phrase embedding and carefully-designed conditions. They mainly assess the credibility of the statistical data explanations and their variants. The three methods all employ semantic relevance degrees as the basis of their judgments. Each relevance degree is either a semantic similarity between a pair of phrases or a ratio of such semantic similarities. The semantic similarity $\mathrm{Sim}(\cdot)$ is a cosine-similarity of the embedded vectors of the phrases generated by Sentence-BERT [63], which is a state-of-the-art deep model for sentence and phrase embeddings. Specifically, the semantic similarity $\mathrm{Sim}(\cdot)$ between $X$ and $Y$ is given as follows.

$$\mathrm{Sim}(X,Y) = \frac{s(X)\cdot s(Y)}{\left\|s(X)\right\|\left\|s(Y)\right\|}, \tag{3.1}$$

where $s(\cdot)$ represents the function of Sentence-BERT to generate embedding vectors.

### 3.5.1 Judgment Method $\alpha$

Judgment method $\alpha$ is devised to judge explanations in $(\alpha)$ category, which take the form of a habit $X$ and a disease $Y$. This method judges an explanation as credible and unethical based on three conditions: 1) if habit $X$ is bad, 2) if disease $Y$ is dangerous, and 3) if the two are highly relevant. These three conditions correspond to three types

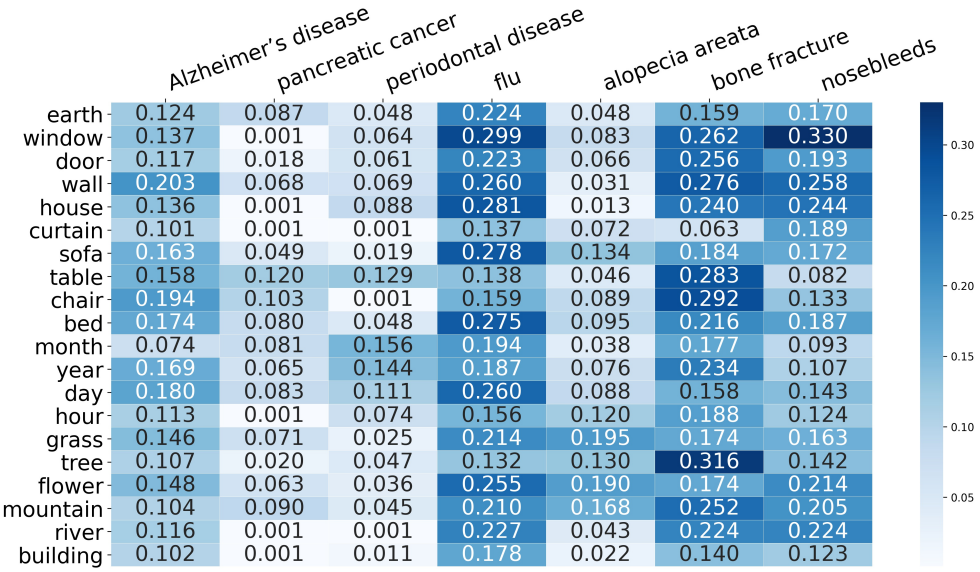



Figure 3.5: Relevance degrees between diseases ($Y$) and base words for $X$ for method $\alpha$.

of credibility.

$$\text{IF } (\theta_{\text{relevance}} > \theta_1) \wedge (\theta_{\text{fear}} > \theta_2) \wedge (\theta_{\text{bad habit}} > \theta_3), \text{ THEN } 1, \text{ ELSE } 0, \tag{3.2}$$

where $\theta_1, \theta_2, \theta_3$ are user-supplied thresholds and $\theta_{\text{relevance}}$, $\theta_{\text{fear}}$, $\theta_{\text{bad habit}}$ are the relevance ratio, the fear ratio, and the bad habit ratio, respectively, given below.

$\theta_{\text{relevance}}$ represents the relevance between $X$ and $Y$. The values of the semantic similarity $\text{Sim}(\cdot)$ largely depend on its arguments, i.e., $X$ and $Y$, which forbids to use the same value for $\theta_1$ in different explanations. To mitigate this variability, $\theta_{\text{relevance}}$ is defined as follows, by selecting a base word of which semantic similarity is neutral to major diseases.

$$\theta_{\text{relevance}} = \frac{\text{Sim}(X, Y)}{\text{Sim}(\text{base word}, Y)}. \tag{3.3}$$

We conducted preliminary experiments on the relevance degrees between a disease and such words. The results are shown in Figure 3.5, where each row and column

represent a candidate of the base word and a major disease[1], respectively. We see that "day" and "earth" have neither small nor large degrees to all the tested diseases. We select the latter as the base word. Similar relevance degrees using base words can be found in a sentiment analysis paper [120]. However, Araque et al. adopted the maximum value for a set of base words for each lexicon word for normalization [120]. On the other hand, we select a base word of which value is not extreme for all diseases to use in Equation 3.3. This difference is due to the nature of the two target problems.

$\theta_{\text{bad habit}}$ represents the bad habit ratio of $X$.

$$\theta_{\text{bad habit}} = \frac{\text{Sim}(X,\ \text{“bad habit”})}{\text{Sim}(X,\ \text{“good habit”})}. \tag{3.4}$$

Note that $\theta_{\text{bad habit}}$ compares the closeness of the habit in $X$ to the phrase "bad habit" than "good habit" through the ratio of the two semantic similarities.

Finally, $\theta_{\text{fear}}$ represents the fear ratio of $Y$. $\theta_{\text{fear}}$ was first devised as follows.

$$\theta_{\text{fear}} = \frac{\text{Sim}(Y,\ \text{“major illness”})}{\text{Sim}(Y,\ \text{“minor illness”})}. \tag{3.5}$$

However, a series of preliminary experiments proved that Equation 3.5 shows quite counter-intuitive results, probably due to our highly-variable subjectivity in assessing the risk of diseases[2]. To address this issue, we adopt a summary (GBD Cause and Risk Summaries[3]) of Disability Adjusted Life Years (DALYs) in Burden of Disease[4] as $\theta_{\text{fear}}$. DALYs are a time-based measure that combines years of life lost due to premature mortality (YLLs) and years of life lost due to time lived in states of less

[1]The diseases are sorted from the heaviest one to the lightest one from the leftmost to the rightmost with a method explained later.

[2]For instance, we have witnessed a young man and a middle-aged man having very different opinions on alopecia areata.

[3]https://www.thelancet.com/gbd/summaries

[4]https://ourworldindata.org/burden-of-disease

than full health, or years of healthy life lost due to disability (YLDs). One DALY represents the loss of the equivalent of one year of full health. DALYs are calculated by the sum of YLLs and YLDs, which allows us to compare different diseases and other kinds of damages.

Take type I explanation "Deep-fried food boosts pancreatic cancer risk" as an example. $X$: "deep-fried food" and $Y$: "pancreatic cancer" are specified as input. If these phrases satisfy the conditions in Equation 3.2, i.e., "deep-fried food" being a bad habit, "pancreatic cancer" being a dangerous disease, and the two being highly relevant, method $\alpha$ judges this explanation as class 1, i.e., credible and unethical.

### 3.5.2 Judgment Method $\beta$

Judgment method $\beta$ is devised to judge explanations in $(\beta)$ category, which take the form of subject $X$ being more likely to have property $Y$ compared with other subjects. To judge the explanation, 5 kinds of phrases $X$, $X'$, $Y_{\text{base}}$, $Y$, and $\overline{Y}$ are specified. $X$ and $Y$ are explicitly mentioned in the explanation. $X'$ is a subject or a set of subjects in the opposite class of $X$, which can be specified explicitly or generated based on knowledge of the English language. $\overline{Y}$ is specified as the inverse property of $Y$, which is typically in the form of an adjective followed by a noun phrase $Y_{\text{base}}$.

Method $\beta$ judges the explanation based on two conditions: 1) if subject $X$ is more relevant to property $Y$ than its inverse property $\overline{Y}$ and 2) if property $Y$ is more relevant to subject $X$ than any other subject $X'$ belonging to the opposite class.

$$\text{IF } (\theta^{\beta}_{XY} > \theta^{\beta}_{X\overline{Y}}) \wedge \forall X'(\theta^{\beta}_{XY} > \theta^{\beta}_{X'Y}), \text{ THEN 1, ELSE 0,} \tag{3.6}$$

where $\theta^{\beta}_{XY}$, $\theta^{\beta}_{X\overline{Y}}$, and $\theta^{\beta}_{X'Y}$ represent the semantic relevance degrees between $X$ and $Y$, $X$ and $\overline{Y}$, as well as $X'$ and $Y$, respectively. The relevance degrees are computed by the

semantic similarities between phrases as follows.

$$\theta^{\beta}_{XY} = \frac{\mathrm{Sim}(X,Y)}{\mathrm{Sim}(X,Y_{\mathrm{base}})},\ \theta^{\beta}_{X\overline{Y}} = \frac{\mathrm{Sim}(X,\overline{Y})}{\mathrm{Sim}(X,Y_{\mathrm{base}})},\ \theta^{\beta}_{X'Y} = \frac{\mathrm{Sim}(X',Y)}{\mathrm{Sim}(X',Y_{\mathrm{base}})}. \tag{3.7}$$

Note that the variability problem of the semantic similarity $\mathrm{Sim}(\cdot)$ also exists in method $\beta$ and the meaning of an adjective is decided by the following noun phrase. Thus the noun phrase $Y_{\mathrm{base}}$ is selected to mitigate this problem, which serves as a base in comparison.

Take type IV explanation "Muslims have many babies compared to Christians" as an example. $X$ and $X'$ are "Muslims" and "Christians", respectively. $Y$ and $\overline{Y}$ are "many babies" and "few babies" with respect to a base word $Y_{\mathrm{base}}$, i.e., "babies", respectively. If these phrases satisfy the conditions in Equation 3.6, i.e., "Muslims" is more relevant to "many babies" than "few babies" and "many babies" is more relevant to "Muslims" than "Christians", method $\beta$ judges this explanation as credible and unethical. However, the semantic similarity between "Muslims" and "many babies" $\mathrm{Sim}(\text{"Muslims"},\text{"many babies"}) = 0.234$ is counter-intuitively lower than the similarity between "Muslims" and "few babies" $\mathrm{Sim}(\text{"Muslims"},\text{"few babies"}) = 0.245$. Thus the first condition is not satisfied, leading to a false negative.

It should be noted that type II (including II-1 and II-2) explanations have no $Y_{\mathrm{base}}$ because their properties, i.e., "poorest" and "richest", take the form of the superlative of an adjective. In such a case, the semantic relevance degrees between phrases are directly calculated by their semantic similarities without considering $Y_{\mathrm{base}}$, e.g., $\theta^{\beta}_{XY} = \mathrm{Sim}(X,Y)$.

### 3.5.3 Judgment Method γ

Judgment method γ is devised to judge explanations in (γ) category, which take the form of subject $X$ having a trend $Y$. To judge the explanation, 3 kinds of phrases $X$, $Y$, and $Y'$ are specified. $X$ and $Y$ are explicitly mentioned in the explanation. $Y'$ is a set of other kinds of trends. Similar to specifying $X'$, $\overline{Y}$, and $Y_{\text{base}}$ for explanations in (β) category, $Y'$ is also generated based on knowledge of the English language.

Method γ judges the explanation based on the condition that if subject $X$ is more relevant to trend $Y$ than any different trend $Y'$. The condition is implemented by comparing if relevance degree $\theta^{\gamma}_{XY}$ between subject $X$ and trend $Y$ is larger than relevance degree $\theta^{\gamma}_{XY'}$ between $X$ and any other relevant trend $Y'$.

$$\text{IF } \forall Y'(\theta^{\gamma}_{XY} > \theta^{\gamma}_{XY'}), \text{ THEN 1, ELSE 0.} \tag{3.8}$$

We typically specify the relevant trends in $Y'$ by replacing the verb or the adverb used in $Y$ with the opposite one, keeping other words as they are[1]. Unlike methods α and β, the variability problem is not serious due to the forms of $Y$ and $Y'$ in (γ) category. Therefore, $\theta^{\gamma}_{XY}$ and $\theta^{\gamma}_{XY'}$ are directly defined as the semantic similarities between the phrases as follows.

$$\theta^{\gamma}_{XY} = \text{Sim}(X,Y),\ \theta^{\gamma}_{XY'} = \text{Sim}(X,Y'). \tag{3.9}$$

Take type XIII explanation "The risk of death from cancer is increasing worldwide" as an example. $X$: "risk of death from cancer", $Y$: "increasing", and $Y' \in$ {"decreasing", "constant"} are specified as input. If these phrases satisfy the conditions in Equation 3.8, i.e., "risk of death from cancer" is more relevant to "increasing" compared with "decreasing" and "constant", respectively, method γ judges this expla-

[1]We used the term "typically" as we also replace a specific trend with a neutral trend. See the example of type XIII.

Table 3.1: Results of methods α, β, and γ.

| α | Predicted Positive | Predicted Negative | β | Predicted Positive | Predicted Negative | γ | Predicted Positive | Predicted Negative |
|---|---|---|---|---|---|---|---|---|
| Actual Positive | 9 | 0 | Actual Positive | 12 | 32 | Actual Positive | 13 | 4 |
| Actual Negative | 0 | 19 | Actual Negative | 13 | 35 | Actual Negative | 3 | 62 |

nation as credible and unethical.

### 3.5.4 Experiments

We conduct experiments on 18 types (I-XVIII)[1] of statistical data explanations to evaluate the judgment methods α, β, and γ. We choose a Sentence-BERT model named "all-mpnet-base-v2"[2] trained on a large amount of data (more than 1 billion training pairs) which can map each phrase to a 768 dimensional dense vector. The thresholds $\theta_1$, $\theta_2$, and $\theta_3$ are all set to 1 in method α.

Table 3.1 shows the confusion matrices of the three judgment methods. The accuracies of the methods α, β, and γ on the 18 types of explanations are 1, 0.511, and 0.918, respectively. We see that methods α and γ exhibit relatively high accuracies in judging the explanations of (α) and (γ) categories, respectively, probably due to the simpler nature of their target explanations. Though method β needs substantial refinement, we believe that the results are quite promising as the first step toward judging credible and unethical explanations of statistical data.

We show detailed results of the three methods in Appendix B. In summary, our method α exhibits the perfect results, which proves the effectiveness of our proposed

[1]Types XIX-XXI were not employed as they were invented later.
[2]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

ratios $\theta_{\text{relevance}}$, $\theta_{\text{fear}}$, and $\theta_{\text{bad habit}}$ for representing the semantic relevance between diseases and habits. Likewise, our method $\gamma$ exhibits an accuracy of nearly 92%. Most of the mistakes (6 out of 7) are committed on type VIII explanation, possibly due to the difficulty in handling a negated phrase, i.e., "not decreasing". With a hindsight, we see that changing "not decreasing" in $Y$ from class label 1 to 0 reduces 3 mistakes and "increasing" from class label 0 to 1 reduces 1 mistake. We do not adopt such a modification but learn the difficulty in handling these expressions and capturing the credibility of people. In overall, however, our method $\gamma$ achieves impressive results of nearly 92%, which proves the effectiveness of our approach.

On the contrary, the performance of our method $\beta$ when judging $(\beta)$ category seems to be unacceptable because it is below the accuracy (0.543) of the baseline method which always predicts the majority class label 0. A closer look at the performance of each type reveals a binary classification of the 11 types. The first group consists of easier types, i.e., type V (7-1), type XIV (7-1), type II-1 (5-3), type VI (5-3), type XVII (5-3), and type XII (4-2), where the numbers in parentheses represent correct and wrong predictions in this order. The accuracy of method $\beta$ is 0.717 (33/46), which we believe is relatively high due to the difficulty of the target problems. On the other hand, the second group consists of more difficult types, i.e., type IV (3-3), type XVIII (4-6), type II-2 (3-5), type XVI (1-9), and type XV (0-4). The accuracy of method $\beta$ is 0.304 (14/46).

As we explained in Chapter 3.4, type XVIII has a serious flaw, i.e., the difficulty of discussing the remaining diseases all at once, but was kept to show the difficulty of the problem. Type XVI also shows the difficulty in handling a serious issue related to the recent pandemic, which hasn't been clarified scientifically and is a subject of fierce debate. Type XV also poses a challenge because small hospitals are in general less well-equipped but receive fewer serious patients than large hospitals. Their degrees of

safety are controversial, which might have influenced our phrase embedding. Omitting these three types yields an accuracy of 0.617 (42/68) for the remaining types, which we believe promising.

A closer look at the remaining difficult types revealed challenges in our defined semantic similarity and relevance degree. Most false predictions are due to the counter-intuitive semantic similarities between the specified phrases for judgment. In type II-2, 4 out of the 5 mistakes were due to the fact that Sim("Somalia","richest") = 0.29137 being larger than that for "UAE" (0.29136), "Qatar" (0.244), "Equatorial Guinea" (0.238), and "Botswana" (0.271). These results are against the statistical data of type II shown in Figure 3.2. The remaining mistake was caused by our choice of "Japan" and "Singapore" for $X'$ in "Botswana is the poorest of the healthiest countries" (class 0). The semantic similarities with "poorest" are 0.179, 0.162, and 0.278 for "Japan", "Singapore", and "Botswana", respectively, which match the data of type II in Figure 3.2. However, this explanation was labeled as 0 due to the numerous countries that have longer life expectancy (healthier) and smaller GDP per capita (poorer) than "Botswana". We believe our choice of $X'$ is correct, as the two countries are representatives of the healthiest ones. We thought of comparing "Botswana" with countries of similar life expectancy, but gave it up as they would not be recognized as "the healthiest countries". Similarly, the majority of the mistakes in types IV and VII were due to semantic similarity that is against our class labeling, e.g., "Muslims" are more similar to "few babies" than "many babies", "Christians" are more similar to "many babies" than "few babies". The remaining mistakes were due to the division by the semantic similarity to the base word, e.g., Sim("Judaisms","babies") = 0.278 is small and thus boosts the relevance degree while that for Christians (0.446) and Muslims (0.437) do not. Note that the semantic similarity and relevance degree in method $\beta$ are effective for many explanations, which proves the difficulty in handling these exceptions.

In summary, the examples of counter-intuitive semantic similarities may be attributed to the fact that our phrase embeddings are directly generated from the pre-trained Sentence-BERT. Sentence-BERT utilizes multiple language datasets for training, without fine-tuning on task-specific corpus. Thus there may exist several phrase embeddings that harm our credibility judgment task. In general, the relevance degree defined by semantic similarity exhibits encouraging performance on judging the credibility of the explanations on statistical data. The underlying semantic relatedness between phrases is worth exploring in the next step.

## 3.6 Graph-Based Judgment Method $\beta^2$

As we discussed in Chapter 3.5.4, the experimental results show that methods $\alpha$ and $\gamma$ exhibit nearly perfect performance, respectively, due to the simple nature of their target explanations in ($\alpha$) and ($\gamma$) categories. However, since the phrases used in the conditions for comparison are more complex in ($\beta$) category, including multiple subjects and properties, several counter-intuitive semantic similarities between these subjects and properties lead to undesired results of method $\beta$.

To address this limitation and improve the low accuracy of method $\beta$, we introduce a new graph-based method $\beta^2$. Method $\beta^2$ first considers more phrases by utilizing their synonyms and constructs a Phrase Similarity Graph to model these phrases and their semantic similarities. Afterward, the conditions for judgment are generated from subgraphs selected from the Phrase Similarity Graph. The importance of the conditions generated from the subgraphs is quantified by their sub-graph entropy. Lastly, a credibility score is devised by aggregating the conditions with their importance to judge the explanations.

We show the overall procedure of method $\beta^2$ in Algorithm 1. Given an explanation,

**Algorithm 1** Overall procedure of method $\beta^2$.

**Input:** Statistical data explanation, Phrases $X$, $X'$, $Y_{\text{base}}$, $Y$, and $\overline{Y}$, Credibility threshold $\theta_{\text{credible}}$.

**Output:** Credible and unethical (class label 1) or not (class label 0).

1: $\mathcal{X}_{\text{syno}}, \mathcal{X'}_{\text{syno}}, \mathcal{Y}_{\text{base,syno}}, \mathcal{Y}_{\text{syno}}, \overline{\mathcal{Y}}_{\text{syno}} = \text{Extend}(X, X', Y_{\text{base}}, Y, \overline{Y})$;

2: Phrase Similarity Graph $\mathcal{G} = \text{GetGraph}(\mathcal{X}_{\text{syno}}, \mathcal{X'}_{\text{syno}}, \mathcal{Y}_{\text{base,syno}}, \mathcal{Y}_{\text{syno}}, \overline{\mathcal{Y}}_{\text{syno}})$;

3: Subgraphs $\mathcal{G}^k, (k = 1, ..., K) = \text{GetSubgraph}(\mathcal{G})$;

4: **for** $k = 1, ..., K$ **do**

5: Generate conditions via four criteria (1)-(4);

6: Calculate sub-score $s_k$ via Equation 3.17;

7: Calculate sub-graph entropy $H(\mathcal{G}^k)$ via Equation 3.18;

8: **end for**

9: Calculate important weight $\lambda_k$ for each sub-score via Equation 3.20;

10: Calculate credibility score $S$ for the explanation via Equation 3.21;

11: If $S > \theta_{\text{credible}}$, output class label 1; else output class label 0.

in step 1, the specified phrases $X$, $X'$, $Y_{\text{base}}$, $Y$, and $\overline{Y}$ are extended to the phrase sets $\mathcal{X}_{\text{syno}}$, $\mathcal{X'}_{\text{syno}}$, $\mathcal{Y}_{\text{base,syno}}$, $\mathcal{Y}_{\text{syno}}$, and $\overline{\mathcal{Y}}_{\text{syno}}$, respectively. The Phrase Similarity Graph $\mathcal{G}$ is constructed based on the extended phrase sets in step 2 and the subgraphs $\mathcal{G}^k$ are extracted by selecting node groups from $\mathcal{G}$ in step 3. Then the conditions for judgment are generated via four criteria (1)-(4) based on the selected node combinations in the subgraph $\mathcal{G}^k$. We are going to explain the details of each step in the rest of this Chapter.

### 3.6.1 Phrase Similarity Graph for Statistical Data Explanations

Given a statistical data explanation with its phrases, we first generate more phrases by considering their synonyms. Then we construct a Phrase Similarity Graph to model an explanation by representing its phrases as nodes and the semantic similarities between different sets of nodes as edges.

Since the unsatisfactory performance of method β is due to the counter-intuitive semantic similarities between limited phrases, we propose to consider more phrases to explore their relevance for judgment. Specifically, each kind of phrase is extended to a phrase set by considering its synonyms. As explained in Chapter 3.5.2, we specify 5 kinds of phrases for judging each explanation in $(\beta)$ category, i.e., $X$, $X'$, $Y_{\text{base}}$, $Y$, and $\overline{Y}$. To obtain more phrases, we adopt an emerging powerful large language model ChatGPT[1] to generate top-$n$ synonyms of each phrase, as we will introduce the details in Chapter 3.6.6.2. The extended phrase sets are represented as $\mathcal{X}_{\text{syno}}$, $\mathcal{X}'_{\text{syno}}$, $\mathcal{Y}_{\text{base,syno}}$, $\mathcal{Y}_{\text{syno}}$, and $\overline{\mathcal{Y}}_{\text{syno}}$ according to $X$, $X'$, $Y_{\text{base}}$, $Y$, and $\overline{Y}$, respectively.

We propose a Phrase Similarity Graph to explicitly model the phrase sets and their semantic similarities. Following several graph-based works [121–123], our graph is an attributed graph defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{W})$, where $\mathcal{V} = \{v_1, \ldots, v_n\}$ represents the set of nodes. $\mathbf{X} \in \mathbb{R}^{n \times d}$ represents the attribute matrix, where the vector $\mathbf{x}_i \in \mathbb{R}^d$ in $\mathbf{X}$ represents the attribute of node $v_i$. $\mathcal{E} = \{e_{i,j} | i, j = 1, \ldots, n\}$ and $\mathbf{W} = \{w_{v_i,v_j} | i, j = 1, \ldots, n\}$ represent the set of edges with weights between nodes $v_i$ and $v_j$, respectively.

In the Phrase Similarity Graph, a node, a node attribute, and an edge with the weight between two nodes represent a phrase, a phrase embedding vector, and a semantic relation computed by cosine-similarity between two phrases in the explanation, respectively. As shown in Figure 3.6, the graph is constructed as a tripartite graph

[1] https://openai.com/blog/chatgpt

$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{W})$ where node set $\mathcal{V}$ consists of three disjoint node subsets $\mathcal{V}_{\text{base}}$, $\mathcal{V}_{\text{subject}}$, and $\mathcal{V}_{\text{property}}$. Nodes in $\mathcal{V}_{\text{base}}$ represent the phrases of base words in $\mathcal{Y}_{\text{base,syno}}$. Nodes in $\mathcal{V}_{\text{subject}}$ represent the phrases of subjects in $\mathcal{X}_{\text{syno}} \cup \mathcal{X}'_{\text{syno}}$. Nodes in $\mathcal{V}_{\text{property}}$ represent the phrases of properties in $\mathcal{Y}_{\text{syno}} \cup \overline{\mathcal{Y}}_{\text{syno}}$.

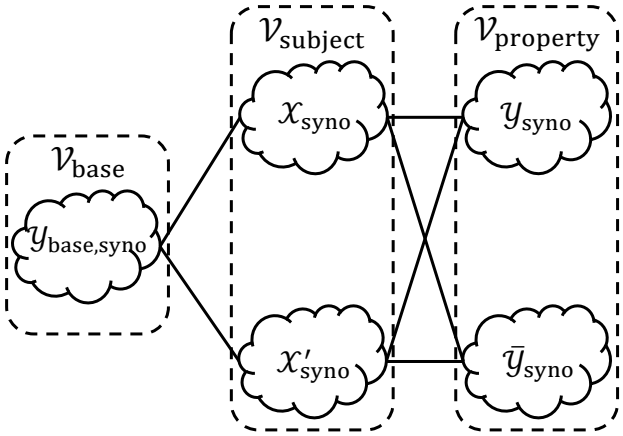



Figure 3.6: Phrase Similarity Graph to model phrase sets and their semantic similarities.

Following method β, we consider the semantic similarities between different kinds of phrases for judgment, i.e., the similarities between subjects and base words and the similarities between subjects and properties. Therefore, the edges in $\mathcal{E}$ are built between nodes in $\mathcal{V}_{\text{base}}$ and $\mathcal{V}_{\text{subject}}$, as well as between nodes in $\mathcal{V}_{\text{subject}}$ and $\mathcal{V}_{\text{property}}$, respectively. The node attribute matrix $\mathbf{X}$ is composed of embedding vectors of phrases, which are generated by Sentence-BERT [63]. The edge weight $w_{v_i,v_j}$ in $\mathbf{W}$ represents the semantic similarity between two nodes $v_i$ and $v_j$, which is calculated by the cosine-similarity $\text{Sim}(\cdot)$ between their node attributes $\mathbf{x}_i$ and $\mathbf{x}_j$. Formally, the edge weight $w_{v_i,v_j}$ between nodes $v_i$ and $v_j$ is given as follows.

$$w_{v_i,v_j} = \text{Sim}(v_i, v_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|}. \tag{3.10}$$

We present an example of a Phrase Similarity Graph for type XII statistical data explanation in Figure 3.7. Given the explanation, each kind of phrase is first extended as a phrase set by considering its synonyms. Then the Phrase Similarity Graph is constructed to represent all the phrases in the phrase sets as nodes and the semantic similarities between nodes from different subsets of nodes as edges. For simplicity, heavy edge weights are shown by the thick width of edges in the graph in Figure 3.7.

(Type XII) Asia is the cause of the large amount of CO2 emissions.
$X$: Asia, $Y$: large amount of CO2 emissions.
(Phrases) $X$: Asia, $X' \in$\{Africa, Europe\}, $Y_{\text{base}}$: CO2 emissions, $Y$: large amount of CO2 emissions, $\overline{Y}$: small amount of CO2 emissions.
(Phrase sets) $\mathcal{X}_{\text{syno}}$:\{Asia, Asian countries, Asian nations\},
$\mathcal{X}'_{\text{syno}}$:\{Europe, European countries, European nations, Africa, African countries, African nations\},
$\mathcal{Y}_{\text{base,syno}}$:\{CO2 emissions, greenhouse gas emissions, carbon dioxide emissions\},
$\mathcal{Y}_{\text{syno}}$:\{large amount of CO2 emissions, large amount of greenhouse gas emissions, large amount of carbon dioxide emissions\},
$\overline{\mathcal{Y}}_{\text{syno}}$:\{small amount of CO2 emissions, small amount of greenhouse gas emissions, small amount of carbon dioxide emissions\}.

### 3.6.2 Additional Conditions by Subgraphs

In addition to the two designed comparison conditions in method β, we propose to consider further conditions for judgment. In method β, the two conditions are devised to compare the relevance between subject $X$ and different properties $Y$ and $\overline{Y}$, as well as property $Y$ with different subjects $X$ and $X'$, represented as $\theta^{\beta}_{XY} > \theta^{\beta}_{X\overline{Y}}$ and $\theta^{\beta}_{XY} > \theta^{\beta}_{X'Y}$.

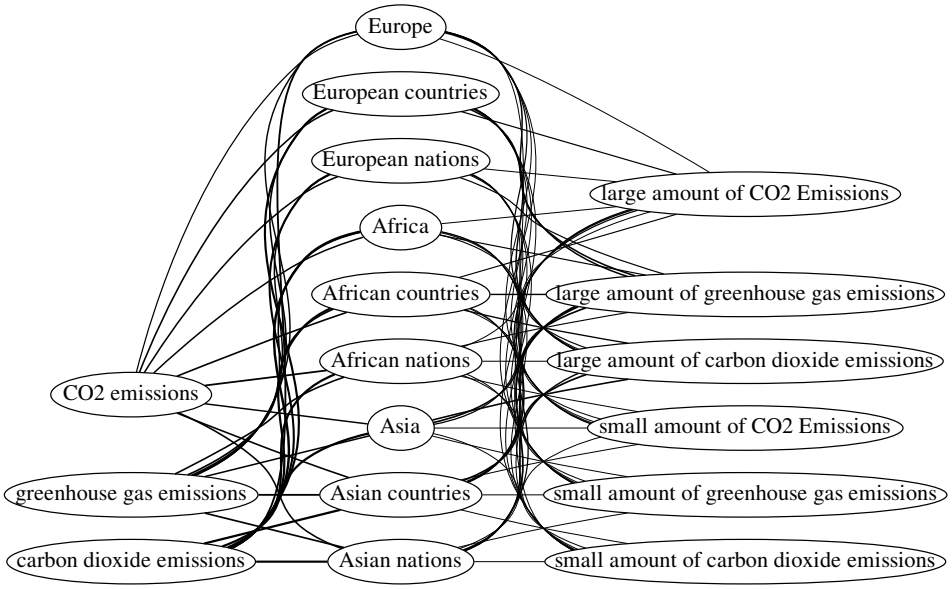



Figure 3.7: Constructing the Phrase Similarity Graph for type XII statistical data explanation. The graph considers 2 synonyms for each kind of phrase.

However, the relevance between the opposite subjects $X'$ with different properties $Y$ and $\overline{Y}$, as well as the opposite property $\overline{Y}$ with different subjects $X$ and $X'$, represented as $\theta^{\beta}_{X'\overline{Y}} > \theta^{\beta}_{X'Y}$ and $\theta^{\beta}_{X'\overline{Y}} > \theta^{\beta}_{X\overline{Y}}$, has not been considered, though these conditions also have the potential to contribute to the judgment. Nevertheless, as shown in Figure 3.7, as the number of phrases increases in the extended phrase sets, designing necessary conditions for judgment becomes more difficult and complex. To simplify the design of conditions, we propose to generate the necessary conditions from the subgraphs extracted from the Phrase Similarity Graph.

Specifically, each subgraph is extracted by selecting one node from each phrase set, e.g., $X$, $X'$, $Y_{\text{base}}$, $Y$, and $\overline{Y}$ from $\mathcal{X}_{\text{syno}}$, $\mathcal{X}'_{\text{syno}}$, $\mathcal{Y}_{\text{base,syno}}$, $\mathcal{Y}_{\text{syno}}$, and $\overline{\mathcal{Y}}_{\text{syno}}$ with the weighted edges, represented as $\mathcal{G}^k = (\mathcal{V}^k, \mathcal{E}^k, \mathbf{X}^k, \mathbf{W}^k), k = 1, \ldots, K$, where $K$ is the number of all subgraphs extracted from Phrase Similarity Graph $\mathcal{G}$. As the conditions

are to compare the relevance between subjects and properties for judgment, the subjects in the opposite classes, i.e., nodes in $\mathcal{X}_{\text{syno}}$ and $\mathcal{X}'_{\text{syno}}$, are selected in pairs. Similarly, two opposite properties, i.e., nodes in $\mathcal{Y}_{\text{syno}}$ and $\overline{\mathcal{Y}}_{\text{syno}}$, are also selected in pairs with the same base word in $\mathcal{Y}_{\text{base,syno}}$. Take the phrase sets in type XII explanation as an example, The subject "Asian countries" is selected together with the other subjects "African countries" and "European countries". Similarly, the properties "large amount of CO2 emissions" and "small among of CO2 emissions" are selected together with the base word "CO2 emissions". Therefore, when extracting a subgraph from the Phrase Similarity Graph, the nodes from $\mathcal{X}_{\text{syno}}$ and $\mathcal{X}'_{\text{syno}}$, as well as the nodes from $\mathcal{Y}_{\text{base,syno}}$, $\mathcal{Y}_{\text{syno}}$, and $\overline{\mathcal{Y}}_{\text{syno}}$ are selected in pairs to generate conditions for judgment.

As each subgraph $\mathcal{G}^k$ represents a group of subjects and properties with their semantic similarities, the aforementioned conditions from a subgraph can be simply generated by considering the node combinations constructed by each node and its neighboring nodes in $\mathcal{V}_{\text{subject}} \cup \mathcal{V}_{\text{property}}$. Given the node combinations, we design the conditions by comparing the relevance between the nodes of subjects and the nodes of properties. The conditions for judgment are designed following four criteria.

(1) Nodes in $\mathcal{X}_{\text{syno}}$ are more relevant to nodes in $\mathcal{Y}_{\text{syno}}$ than nodes in $\overline{\mathcal{Y}}_{\text{syno}}$.

(2) Nodes in $\mathcal{X}'_{\text{syno}}$ are more relevant to nodes in $\overline{\mathcal{Y}}_{\text{syno}}$ than nodes in $\mathcal{Y}_{\text{syno}}$.

(3) Nodes in $\mathcal{Y}_{\text{syno}}$ are more relevant to nodes in $\mathcal{X}_{\text{syno}}$ than nodes in $\mathcal{X}'_{\text{syno}}$.

(4) Nodes in $\overline{\mathcal{Y}}_{\text{syno}}$ are more relevant to nodes in $\mathcal{X}'_{\text{syno}}$ than nodes in $\mathcal{X}_{\text{syno}}$.

Following the definition in Chapter 3.5.2, we adopt $\theta^{\beta}_{XY}$ as the semantic relevance degree between two nodes $X$ and $Y$, which can be calculated by the edge weights in our graph.

$$\theta^{\beta}_{XY} = \frac{\text{Sim}(X,Y)}{\text{Sim}(X,Y_{\text{base}})} = \frac{w_{X,Y}}{w_{X,Y_{\text{base}}}}. \tag{3.11}$$

We present a subgraph of type XII explanation in Figure 3.8 extracted from the Phrase Similarity Graph in Figure 3.7 by selecting a group of nodes $X$, $X'_1$, $X'_2$, $Y_{\text{base}}$, $Y$, and $\overline{Y}$. The numbers in the subgraph represent the edge weights, i.e., the semantic similarities, between nodes.

(Type XII) Asia is the cause of the large amount of CO2 emissions.
(Selected nodes from Phrase Similarity Graph) $X$: Asian countries, $X'_1$: African countries, $X'_2$: European countries, $Y_{\text{base}}$: CO2 emissions, $Y$: large amount of CO2 emissions, $\overline{Y}$: small amount of CO2 emissions.

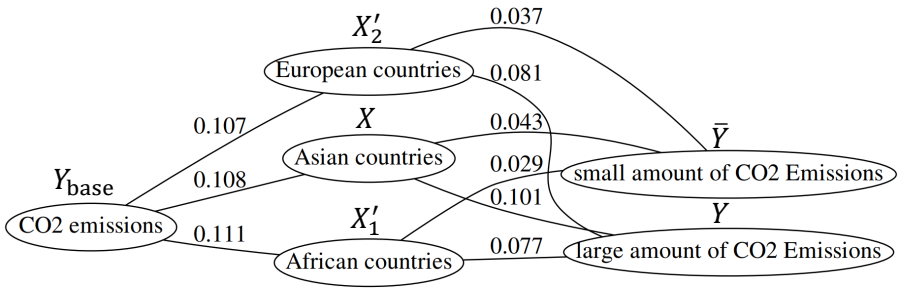



Figure 3.8: Example of a subgraph extracted from the Phrase Similarity Graph of type XII explanation.

In the subgraph, by selecting each node and its neighboring nodes in $\mathcal{V}_{\text{subject}} \cup \mathcal{V}_{\text{property}}$, the conditions are generated by following the four criteria (1)-(4) as follows.

$$X \cup \mathcal{N}(X) = \{X, Y, \overline{Y}\} \rightarrow \text{IF } \theta^{\beta}_{XY} > \theta^{\beta}_{X\overline{Y}}, \tag{3.12}$$

$$X'_1 \cup \mathcal{N}(X'_1) = \{X'_1, Y, \overline{Y}\} \rightarrow \text{IF } \theta^{\beta}_{X'_1\overline{Y}} > \theta^{\beta}_{X'_1Y}, \tag{3.13}$$

$$X'_2 \cup \mathcal{N}(X'_2) = \{X'_2, Y, \overline{Y}\} \rightarrow \text{IF } \theta^{\beta}_{X'_2\overline{Y}} > \theta^{\beta}_{X'_2Y}, \tag{3.14}$$

$$Y \cup \mathcal{N}(Y) = \{Y, X, X'_1, X'_2\} \rightarrow \begin{cases} \text{IF } \theta^{\beta}_{XY} > \theta^{\beta}_{X'_1Y}, \\ \text{IF } \theta^{\beta}_{XY} > \theta^{\beta}_{X'_2Y}, \end{cases} \tag{3.15}$$

$$\overline{Y} \cup \mathcal{N}(\overline{Y}) = \{\overline{Y}, X, X_1', X_2'\} \rightarrow \begin{cases} \text{IF } \theta^{\beta}_{X_1'\overline{Y}} > \theta^{\beta}_{X\overline{Y}}, \\ \text{IF } \theta^{\beta}_{X_2'\overline{Y}} > \theta^{\beta}_{X\overline{Y}}, \end{cases} \tag{3.16}$$

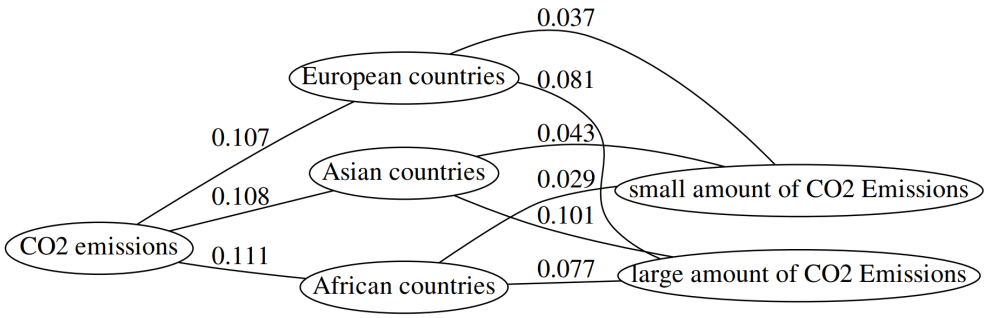
where $\mathcal{N}(X)$ represents the neighboring nodes of node $X$. Among the group of conditions from the subgraph, each satisfied condition increases the credibility of the explanation. We define a sub-score $s_k$ to represent the proportion of satisfied conditions over all conditions from each subgraph $\mathcal{G}^k$ as follows.

$$s_k = \frac{\text{the number of satisfied conditions in } \mathcal{G}^k}{\text{the number of all conditions in } \mathcal{G}^k}. \tag{3.17}$$

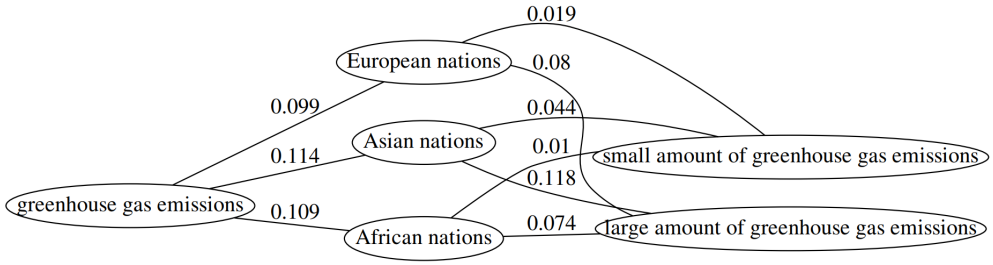
### 3.6.3 Graph Entropy for Importance of Conditions

As we mentioned above, a group of conditions for judging an explanation is generated from each subgraph. Since the conditions for judgment are based on the diverse node attributes and edge weights from different subgraphs, they should be assigned different importance to judge an explanation. For example, Figure 3.9 shows two subgraphs $\mathcal{G}^1$ and $\mathcal{G}^2$ extracted from the Phrase Similarity Graph in Figure 3.7. As the nodes and the edge weights representing their semantic similarities are different in the two subgraphs, the semantic similarity-based conditions generated from $\mathcal{G}^1$ and $\mathcal{G}^2$ should have different importance for judgment.

Sen et al. [119] utilize the sub-graph entropy based on edge weights to calculate the importance of subgraphs in functional brain networks. Following this work, we adopt the sub-graph entropy to quantify the importance of the generated conditions from each subgraph. In our approach, the edge weights refer to the semantic similarities between nodes, so the graph entropy measures the uncertainty of semantic similarities between nodes in a subgraph. A subgraph with high graph entropy indicates a greater uncer-




(a) Subgraph $\mathcal{G}^1$.




(b) Subgraph $\mathcal{G}^2$.

Figure 3.9: Two subgraphs extracted from the Phrase Similarity Graph in Figure 3.7.

tainty in the semantic similarities between its nodes, which suggests that the conditions generated from this subgraph should be assigned less importance. The graph entropy $H(\mathcal{G}^k)$ of a subgraph $\mathcal{G}^k$ is negatively related to the importance of its generated conditions. To keep the weight value within the range of $[0,1]$, we utilize the normalized exponential function of negative graph entropy $e^{-H(\mathcal{G}^k)}$ as the weight $\lambda_k$ to represent the importance of the conditions from the subgraph $\mathcal{G}^k$.

As introduced in [119], sub-graph entropy represents the uncertainty of a subgraph within a whole graph. Sub-graph entropy is calculated by the normalized edge weights, which allows a fair comparison between subgraphs with different ranges of edge weights. Formally, given a subgraph $\mathcal{G}^k = (\mathcal{V}^k, \mathcal{E}^k, \mathbf{X}^k, \mathbf{W}^k)$, its sub-graph entropy

$H(\mathcal{G}^k)$ is calculated as follows.

$$H(\mathcal{G}^k) = -\sum_{i,j} p^k_{v_i,v_j} \log p^k_{v_i,v_j}, \tag{3.18}$$

$$\text{where } p^k_{v_i,v_j} = \frac{w^k_{v_i,v_j}}{\sum_{i,j} w^k_{v_i,v_j}}. \tag{3.19}$$

Take the two subgraphs $\mathcal{G}^1$ and $\mathcal{G}^2$ in Figure 3.9 as an example. Based on the normalized edge weights between nodes, the sub-graph entropy for $\mathcal{G}^1$ and $\mathcal{G}^2$ are calculated as $H(\mathcal{G}^1) = 3.04$ and $H(\mathcal{G}^2) = 2.93$, which indicates that $\mathcal{G}^2$ has less uncertainty in the semantic similarities between its nodes, and thus the conditions generated from $\mathcal{G}^2$ should be assigned more importance. The weight $\lambda_k$ representing the importance of the conditions generated from subgraph $\mathcal{G}^k$ is calculated by the normalized exponential function of negative sub-graph entropy $e^{-H(\mathcal{G}^k)}$ as follows.

$$\lambda_k = \frac{e^{-H(\mathcal{G}^k)}}{\sum_{k=1}^{K} e^{-H(\mathcal{G}^k)}}. \tag{3.20}$$

### 3.6.4 Credibility Score for Judgment

The credibility score of an explanation is defined by summing up all sub-scores and their corresponding weights, which are determined by the conditions generated from all subgraphs and their importance evaluated by sub-graph entropy. Given sub-scores $s_k$ and weight $\lambda_k$ of all subgraphs $\{\mathcal{G}^k | k = 1, \ldots, K\}$, credibility score $S$ is calculated as follows.

$$S = \sum_{k=1}^{K} \lambda_k s_k. \tag{3.21}$$

Credibility score $S$ ranges from 0 to 1, where a higher score indicates stronger credibility of the explanation. We define user-supplied threshold $\theta_{\text{credible}}$ for our method

$\beta^2$. If $S > \theta_{\text{credible}}$, the explanation is judged as credible and unethical.

As we explained in Chapter 3.5.2, type II explanations have no $Y_{\text{base}}$. Therefore, the Phrase Similarity Graph is constructed as a bipartite graph without subset $\mathcal{V}_{\text{base}}$. Accordingly, semantic relevance $\theta^{\beta}_{XY}$ is also simply defined as $\theta^{\beta}_{XY} = \text{Sim}(X,Y)$. The judging procedure for type II explanations is the same as judging other types, except for the shape of the Phrase Similarity Graph and the extracted subgraphs, as well as the definition of $\theta^{\beta}_{XY}$.

### 3.6.5 Complexity Analysis

We analyze the time complexity of the proposed method $\beta^2$ when judging a statistical data explanation. Given a statistical data explanation, let $m$ be the number of its phrases and we consider $n$ synonyms for each phrase. The number of nodes in the Phrase Similarity Graph is $mn$. By considering the semantic similarities between nodes in different subsets to build edges, the time complexity of constructing a Phrase Similarity Graph is $\mathcal{O}(mn^2)$. We propose to extract the subgraphs in the Phrase Similarity Graph by selecting nodes in groups from subjects and properties, respectively, so the time complexity for the extraction is $\mathcal{O}(n^2)$. For each subgraph, the time complexities for generating conditions and calculating its graph entropy are $\mathcal{O}(m^2)$ and $\mathcal{O}(m)$, respectively. Therefore, the time complexity for judging an explanation based on the graph is $\mathcal{O}(m^2n^2)$. To sum up, the overall time complexity for method $\beta^2$ is $\mathcal{O}(m^2n^2)$. In our experiments, the values of $m$ and $n$ are less than ten and there are hundreds of explanations, which demonstrate that our method $\beta^2$ is fast and efficient for the target problem.

### 3.6.6 Experiments

We conduct experiments to evaluate the performance of the proposed method $\beta^2$. The experimental results are illustrated including a comparison of performance and detailed analysis.

#### 3.6.6.1 Datasets

Our method $\beta^2$ is evaluated on statistical data explanations in $(\beta)$ category. Our defined 21 types of explanations contain 14 types within $(\beta)$ category, including types II, IV-VII, XII, and XIV-XXI. The details of the explanations with their statistical data have been introduced in Chapter 3.4. The 14 types of explanations cover a wide range of topics, which involve health, economy, education, collaboration, religion, and energy. The phrases specified from each explanation, including the subject and the property, are shown in Table 3.3.

The total number of the explanations (including their variants) in $(\beta)$ category is 122, consisting of 59 credible and unethical explanations and 63 not credible and unethical explanations. We settle on an approximate $50-50$ class balance in our experiments as it is the most difficult setting for a classification task. The ratio of the anomalies in the real world can vary. We avoid the problem of an arbitrary ratio of anomalies by this equal distribution setting. The ground-truth class labels of these explanations were manually assigned through a careful and consistent discussion among the authors [18].

#### 3.6.6.2 Experimental Setup

We utilize method $\beta$ as the baseline method to evaluate the performance of the proposed method $\beta^2$. To generate phrase sets, we adopt a large language model, ChatGPT

with the released version named "ChatGPT Jan 9 Version"[1] in 2023, to search for the top-$n$ synonyms of each kind of phrase. Specifically, the top-$n$ synonyms are obtained by utilizing the prompt "what are similar words to ⟨phrase⟩?" and selecting the top-$n$ answers, where ⟨phrase⟩ is replaced by each kind of phrase when searching for its synonyms. In our experiments, by investigating the qualities of the generated synonyms, $n$ is set to 3. We notice that some of the synonyms of the proper nouns generated by ChatGPT are far from their original meanings, e.g., "East Asia" is generated as the synonym for "China". Therefore, we exclude the synonyms for phrases which are proper nouns, including countries and disease names in types II, XIV, XVIII, XX, and XXI. When generating phrase embeddings, we choose a Sentence-BERT model named "all-mpnet-base-v2"[2] pre-trained on a large amount of data (more than 1 billion training pairs), which can map each phrase to a 768 dimensional dense vector. The credibility threshold $\theta_{\text{credible}}$ is set to 0.5.

#### 3.6.6.3 Experimental Results and Analysis

The experimental results were obtained by measuring the agreement between the predicted class labels and the ground-truth class labels. Table 3.2 shows the confusion matrices of our graph-based method $\beta^2$ compared with method $\beta$ on the 14 types of statistical data explanations. Due to the large number of false negatives in the results, method $\beta$ exhibits a relatively low accuracy, which is 0.574. In contrast to method $\beta$, our proposed method $\beta^2$ shows a significant improvement, which achieves an accuracy of 0.811. In summary, our method $\beta^2$ significantly outperforms the baseline method $\beta$ with about 0.237 improvement in accuracy. The results demonstrate the effectiveness of the proposed method $\beta^2$ for the target problem.

[1] https://help.openai.com/en/articles/6825453-chatgpt-release-notes
[2] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Table 3.2: Results of method $\beta^2$ compared with method $\beta$.

| $\beta$ | Predicted Positive | Predicted Negative | $\beta^2$ | Predicted Positive | Predicted Negative |
|---|---|---|---|---|---|
| Actual Positive | 20 | 39 | Actual Positive | 48 | 11 |
| Actual Negative | 13 | 50 | Actual Negative | 12 | 51 |

Similar to the experimental results discussed in Chapter 3.5.4, we show detailed results of method $\beta^2$ on 14 types of explanations (including their variants) in Table 3.3. In the Table, each explanation is represented by subject $X$ with property $Y$ in each row, where property $Y$ in parentheses represents that the explanation belongs to class 0. For example, in type IV, $X$: Muslims and $Y$: many babies form an explanation "Muslims have many babies compared to Christians" with class label 1. By replacing the property $Y$, $X$: Muslims and $Y$: (few babies) form its variant "Muslims have few babies compared to Christians" with class label 0. FP and FN with bold fonts represent that the explanation is judged as a false positive and a false negative, respectively. A blank in the Result column either represents a true positive or a true negative.

Based on the results in Table 3.3, our method $\beta^2$ achieves almost perfect performance on 10 types of explanations, including types II (16-0), IV (6-0), V (7-1), VI (8-0), XII (6-2), XIV (7-1), XVIII (9-1), XIX (6-2), XX (16-0), and XXI (6-0), where the numbers in parentheses represent correct and wrong predictions in this order. The baseline method $\beta$ obtains 31 false predictions on these 10 types, where 26 false predictions are caused by the counter-intuitive semantic similarities between subjects and properties. In contrast, our method $\beta^2$ yields 7 false predictions, with only 5 false pre-

dictions caused by this issue. This fact demonstrates that $\beta^2$ is capable of providing more accurate answers compared with method $\beta$. Take an explanation with class 1 in type VI from Table 3.3 as an example, i.e., "Iranians have many children compared to Americans in the 21st century". Method $\beta$ obtains a false negative due to the counter-intuitive semantic similarities between "Iranians" and "many children" (0.285) and between "Iranians" and "few children" (0.294). While our method $\beta^2$ gives a correct answer to this explanation by considering the synonyms in the phrase sets, e.g., "Iranian nationals", "many babies", and "few babies", which do not have counter-intuitive similarities. The 2 false predictions in types V and XIV are attributed to the fact that both two explanations describing one subject with two opposite properties are assigned class 0. The class labels of these two explanations reflect the subjectivity of persons, which is difficult to be estimated with no mistakes.

On the other hand, our method $\beta^2$ achieves relatively low accuracy on 4 types, including VII (4-4), XV (0-4), XVI (6-4), XVII (4-4), by obtaining 16 false predictions, while method $\beta$ obtains 21 false predictions. There exist several explanations where method $\beta^2$ fails while method $\beta$ succeeds. Take an explanation of class 0 in type XVII from Table 3.3 as an example, i.e., "Europe has lower GDP per capita than other regions". Method $\beta$ gives a correct prediction for it as the similarity between "Europe" and "low GDP" (0.300) is intuitively lower than the similarity between "Europe" and "high GDP" (0.369). On the other hand, our method $\beta^2$ yields a false positive because several synonyms in their phrase sets, e.g., "Europe", "weak economy", and "strong economy", have counter-intuitive similarities. However, it is worth noting that our method $\beta^2$ achieves equal or higher accuracy on 12 types (106 explanations) while lower accuracy on only 2 types (16 explanations) compared with method $\beta$. In addition, type XV poses a challenge because small hospitals are in general less well-equipped but receive fewer serious patients than large hospitals. Their degrees of safety are

controversial, which might have influenced the phrase embeddings. Type XVI shows the difficulty in handling a serious issue related to the recent pandemic, which has not been clarified scientifically and is a subject of fierce debate. Omitting these two controversial types, our method $\beta^2$ can achieve an accuracy of 0.861 compared to 0.639 achieved by method $\beta$. We believe these results show the performance of the two methods more appropriately.

We investigate the issue of the counter-intuitive semantic similarities between phrases in the results of the two methods $\beta$ and $\beta^2$ under scrutiny. Take the type IV explanation, "Muslims have many babies compared to Christians", as an example. Method $\beta$ fails in judging it because the semantic similarity between "Muslims" and "many babies" (0.234) is counter-intuitively lower than the similarity between "Muslims" and "few babies" (0.245). In contrast, our method $\beta^2$ succeeds because the majority of the synonyms of "Muslims" exhibits higher semantic similarities to the synonyms of "many babies" compared to the synonyms of "few babies" in the extended phrase sets. For instance, Sim("Muslims", "many infants")=0.230, Sim("Islam followers", "many babies")=0.173, and Sim("Islam followers", "many kids")=0.195 are higher than Sim("Muslims", "few infants")=0.228, Sim("Islam followers", "few babies")=0.165, and Sim("Islam followers", "few kids")=0.177, respectively. The investigation suggests that the intuitive semantic similarities among the majority of the synonyms mitigate the problem of the counter-intuitive similarities between specified phrases, and thus help our credibility score for accurate judgment.

The results of our method show a characteristic that among the explanations in one type, the credibility scores of two explanations which describe one subject with two opposite properties sum up to 1. For example, in type V, the credibility scores for two explanations "women have lower math scores than men" and "women have higher math scores than men" are 0.527 and 0.473, respectively, where the two scores

sum up to 1. The reason is that the two groups of conditions for judging these two explanations are complementary. When the conditions for judging one explanation hold, the complementary conditions for judging the other explanation, which has the same subject with the opposite property, will not be satisfied. This fact leads to the two explanations having complementary credibility scores, which sum up to 1. We believe that this is a desirable characteristic of our proposed method for the target problem since humans are unlikely to believe that a subject can simultaneously have both a property and its opposite property compared to other subjects. Therefore, the credibility scores among the explanations in one type are consistent with this intuition.

## 3.7 Summary

In this Chapter, we have investigated the exploitation of ten human instincts [2] in statistical data explanations as a first yet important step toward ethical AI. We first defined 21 types of credible and unethical explanations of statistical data with the exploitation of the instincts. Then we introduced three methods $\alpha$, $\beta$, and $\gamma$ based on carefully-designed conditions for judging credible and unethical statistical data explanations. Experiments on the statistical data explanations show the effectiveness of methods $\alpha$ and $\gamma$. However, method $\beta$ achieves relatively low accuracy due to the counter-intuitive semantic similarities between phrases when judging the explanations in $(\beta)$ category. To address the limitation and improve the unsatisfactory performance of method $\beta$, we proposed a graph-based method $\beta^2$. In method $\beta^2$, the Phrase Similarity Graph is constructed to explicitly model the phrases in phrase sets and their semantic similarities, where the phrase sets are generated by considering synonyms of phrases specified from the explanation. The credibility score is devised by combining the conditions generated from the Phrase Similarity Graph with their corresponding importance measured

Table 3.3: Detailed results and credibility scores of method $\beta^2$, where the abbreviations MR, ER, CO2E, UNs, MPW, and PE represent mortality rates, enrollment rates, CO2 emissions, United Nations, mismanaged plastic waste, and plastic emissions, respectively.

| Type | *X* | *Y* | Score | Result |
|---|---|---|---|---|
| II-1 | Cuba | poorest | 0.750 | |
| 8-0 | | (richest) | 0.250 | |
| | Nicaragua | poorest | 1.000 | |
| | | (richest) | 0.000 | |
| | Bangladesh | poorest | 0.644 | |
| | | (richest) | 0.356 | |
| | North Korea | poorest | 0.571 | |
| | | (richest) | 0.429 | |
| II-2 | United Arab Emirates | richest | 0.892 | |
| 8-0 | | (poorest) | 0.108 | |
| | Qatar | richest | 0.679 | |
| | | (poorest) | 0.321 | |
| | Equatorial Guinea | richest | 0.785 | |
| | | (poorest) | 0.215 | |
| | Botswana | richest | 0.536 | |
| | | (poorest) | 0.464 | |
| IV | Muslims | many babies | 0.515 | |
| 6-0 | | (few babies) | 0.485 | |
| | Judaisms | many babies | 0.531 | |
| | | (few babies) | 0.469 | |
| | Christians | few babies | 0.515 | |
| | | (many babies) | 0.485 | |
| V | women | low math score | 0.527 | |
| 7-1 | | (high math score) | 0.473 | |
| | men | high math score | 0.527 | |
| | | (low math score) | 0.473 | |
| | women | (low English score) | 0.395 | |
| | | high English score | 0.605 | |
| | men | (high English score) | 0.395 | |
| | | (low English score) | 0.605 | **FP** |
| VI | Iranians | many children | 0.583 | |
| 8-0 | | (few children) | 0.417 | |
| | Afghans | many children | 0.708 | |
| | | (few children) | 0.292 | |
| | French | few children | 0.434 | |
| | | (many children) | 0.566 | |
| | Americans | few children | 0.391 | |
| | | (many children) | 0.609 | |
| VII | developing countries | high infant MR | 0.468 | **FN** |
| 4-4 | | (low infant MR) | 0.532 | **FP** |
| | advanced countries | low infant MR | 0.468 | **FN** |
| | | (high infant MR) | 0.532 | **FP** |
| | developing countries | low ER | 0.718 | |
| | | (high ER) | 0.282 | |
| | advanced countries | high ER | 0.718 | |
| | | (low ER) | 0.282 | |
| XII | Asia | large amount of CO2E | 0.470 | **FN** |
| 4-2 | | (small amount of CO2E) | 0.530 | **FP** |
| | Africa | small amount of CO2E | 0.571 | |
| | | (large amount of CO2E) | 0.429 | |
| | Europe | small amount of CO2E | 0.613 | |
| | | (large amount of CO2E) | 0.387 | |
| XIV | China | large amount of CO2E | 0.750 | |
| 7-1 | | (small amount of CO2E) | 0.250 | |
| | India | large amount of CO2E | 0.686 | |
| | | (small amount of CO2E) | 0.314 | |
| | United States | large amount of CO2E | 0.573 | |
| | | (small amount of CO2E) | 0.427 | |
| | United Kingdom | (large amount of CO2E) | 0.251 | |
| | | (small amount of CO2E) | 0.749 | **FP** |
| XV | small hospitals | dangerous hospitals | 0.354 | **FN** |
| 0-4 | | (safe hospitals) | 0.646 | **FP** |
| | large hospitals | safe hospitals | 0.354 | **FN** |
| | | (dangerous hospitals) | 0.646 | **FP** |
| XVI | Omicron strain | less dangerous | 0.512 | |
| 6-4 | | (more dangerous) | 0.488 | |
| | Alpha strain | less dangerous | 0.497 | **FN** |
| | | (more dangerous) | 0.503 | **FP** |
| | Beta strain | less dangerous | 0.531 | |
| | | (more dangerous) | 0.469 | |
| | Gamma strain | less dangerous | 0.483 | **FN** |
| | | (more dangerous) | 0.517 | **FP** |
| | Delta strain | more dangerous | 0.512 | |
| | | (less dangerous) | 0.488 | |
| XVII | Africa | low GDP | 0.795 | |
| 4-4 | | (high GDP) | 0.205 | |
| | Asia | high GDP | 0.600 | |
| | | (low GDP) | 0.400 | |
| | Americas | high GDP | 0.481 | **FN** |
| | | (low GDP) | 0.519 | **FP** |
| | Europe | high GDP | 0.424 | **FN** |
| | | (low GDP) | 0.575 | **FP** |
| XVIII | cancer | (long life expectancy) | 0.371 | |
| 9-1 | | short life expectancy | 0.629 | |
| | Alzheimer's disease | (long life expectancy) | 0.500 | |
| | | short life expectancy | 0.500 | **FN** |
| | heart disease | (long life expectancy) | 0.436 | |
| | | short life expectancy | 0.564 | |
| | pneumonia | (long life expectancy) | 0.309 | |
| | | short life expectancy | 0.691 | |
| | periodontal disease | (short life expectancy) | 0.326 | |
| | | long life expectancy | 0.674 | |
| XIX | Americas | many members of the UNs | 0.513 | |
| 6-2 | | (few members of the UNs) | 0.487 | |
| | Europe | many members of the UNs | 0.515 | |
| | | (few members of the UNs) | 0.485 | |
| | Asia | many members of the UNs | 0.438 | **FN** |
| | | (few members of the UNs) | 0.562 | **FP** |
| | Africa | few members of the UNs | 0.530 | |
| | | (many members of the UNs) | 0.470 | |
| XX | India | large amount of MPW | 0.576 | |
| 16-0 | | (small amount of MPW) | 0.424 | |
| | China | large amount of MPW | 0.781 | |
| | | (small amount of MPW) | 0.219 | |
| | United Kingdom | small amount of MPW | 0.856 | |
| | | (large amount of MPW) | 0.144 | |
| | United States | small amount of MPW | 0.536 | |
| | | (large amount of MPW) | 0.464 | |
| | India | large amount of PE | 0.573 | |
| | | (small amount of PE) | 0.427 | |
| | China | large amount of PE | 0.644 | |
| | | (small amount of PE) | 0.356 | |
| | United Kingdom | small amount of PE | 0.713 | |
| | | (large amount of PE) | 0.287 | |
| | United States | small amount of PE | 0.664 | |
| | | (large amount of PE) | 0.336 | |
| XXI | Australia | low fossil fuel consumption | 0.750 | |
| 6-0 | | (high fossil fuel consumption) | 0.250 | |
| | United Kingdom | low fossil fuel consumption | 0.810 | |
| | | (high fossil fuel consumption) | 0.190 | |
| | United States | low fossil fuel consumption | 0.750 | |
| | | (high fossil fuel consumption) | 0.250 | |

by the sub-graph entropy. Experiments on the explanations demonstrate the superiority of the proposed method $\beta^2$ on the target problem compared with the baseline method $\beta$.

# Chapter 4

# Conclusions and Future Work

## 4.1 Conclusions

In this thesis, we focused on learning semantic attributed graphs for judging deviated human activity and understanding. We explored two specific tasks within this area, i.e., image region anomaly detection task in human monitoring and judging credible and unethical explanations of statistical data in Chapters 2 and 3, respectively.

In Chapter 2, we first introduced the diverse anomalies, i.e., single and contextual anomalies, at the region level in human monitoring. In addition to considering the spatial relations, we explored the semantic relations among regions and proposed a Spatial and Semantic Attributed Graph to capture the contexts of regions. In the Spatial and Semantic Attributed Graph, each region with its features is represented as a node with attributes. The edges are built between regions by considering their spatial adjacency and semantic similarities between their captions. Then we devised a tailored graph auto-encoder SSGAE with the adoption of the sum aggregation strategy [26]. SSGAE is trained to reconstruct both the node attributes and the node structures in the graph. The attribute and structure reconstruction errors are utilized in the anomaly score to

estimate the abnormality of regions. Our experiments show that our method SSGAE outperforms other baseline algorithms in terms of the ROC curve and AUC score.

In Chapter 3, we first defined 21 types of credible and unethical statistical data explanations with the exploitation of ten human instincts in Rosling et al. [2]. We introduced three judgment methods $\alpha$, $\beta$, and $\gamma$ by comparing the semantic relevance between phrases specified from the explanations. Experiments on the explanations of statistical data show the effectiveness of methods $\alpha$ and $\gamma$. Nevertheless, method $\beta$ exhibits unsatisfactory performance due to the counter-intuitive semantic similarities between phrases when judging the explanations in $(\beta)$ category. To improve the low accuracy of method $\beta$, we proposed a graph-based method $\beta^2$. Method $\beta^2$ first constructs a Phrase Similarity Graph for more reliable semantic relations between phrases in the explanation. By extracting subgraphs in the Phrase Similarity Graph, necessary comparison conditions are generated for judgment. We adopted the graph entropy to quantify the importance of the generated conditions in each subgraph. Lastly, we devised a credibility score to aggregate the satisfied conditions and their importance for more accurate judgment. The experiments on the 14 types of statistical data explanations demonstrate the superiority of the proposed method $\beta^2$ compared with the baseline method $\beta$ in terms of accuracy. In addition, scrutiny reveals that method $\beta^2$ effectively mitigates the problem of counter-intuitive semantic similarities in method $\beta$.

## 4.2 Future Work

In future work, we will consider more informative attributed graphs for better performance in judging deviated human activity and understanding. For the region anomaly detection task in Chapter 2, we plan to construct a weighted attributed graph, where

the weights can represent the importance of relations among regions. Such a model would promote our future method toward more real-world applications in complex scenarios. Our proposed method $\beta^2$ in Chapter 3 opens a new opportunity to bridge the gap between graph models and judging explanations of statistical data. However, we notice that the significance of the statistical data explanation, which is a necessary condition to define the credible and unethical explanation, has not been taken into account. Our future work will extend the Phrase Similarity Graph by incorporating neutral base words and their semantic relations with other phrases. Based on the extended graph, we plan to develop an objective measure of the significance, which can contribute to a more comprehensive judgment of the credible and unethical statistical data explanations.

# Appendix A. Algorithm of SSGAE for Region Anomaly Detection

**Algorithm 2** Overall procedure of SSGAE (training stage).

**Input:** Graph $\mathcal{G}_{\text{train}}^{k} = \{\mathbf{A}^{k}, \mathbf{X}^{k}\}_{k=1}^{K}$, $\mathcal{G}_{\text{test}}^{k'} = \{\mathbf{A}^{k'}, \mathbf{X}^{k'}\}_{k'=1}^{K'}$; Learnable parameter $\Theta$; Hyper-parameter $\beta$; Number $L$ of the hidden layers in SSGAE; Number $T$ of the training epochs.

**Output:** Anomaly score $s_i^{k'}$ for each node $v_i^{k'}$ via function $f(\cdot)$.

▷ Training Stage.
Randomly initialize $\Theta$ and the trainable parameters in $\text{MLP}_{\text{Enc}}$, $\text{MLP}_{\text{Str-Dec}}$ and $\text{MLP}_{\text{Att-Dec}}$;
**for** $t = 1, 2, \cdots, T$ **do**
    **for** $k = 1, 2, \cdots, K$ **do**
        **for** $l = 1, 2, \cdots, L$ **do**
            Calculate $\mathbf{H}^{(l)}$ via Equation 2.3;
        **end for**
        $\mathbf{Z}^{k} = \mathbf{H}^{(L)}$;
        **for** $l = 1, 2, \cdots, L$ **do**
            Calculate $\widehat{\mathbf{H}}^{(l)}$ via Equation 2.7;
        **end for**
        $\widehat{\mathbf{X}}^{k} = \widehat{\mathbf{H}}^{(L)}$;
        Calculate $\widehat{\mathbf{A}}^{k}$ via Equation 2.5;
        Update $\Theta$ and the trainable parameters in $\text{MLP}_{\text{Enc}}$, $\text{MLP}_{\text{Str-Dec}}$, and $\text{MLP}_{\text{Att-Dec}}$ via Equation 2.8 with the backpropagation algorithm.
    **end for**
**end for**

**Algorithm 2** Overall procedure of SSGAE (test stage).

▷ Test Stage.
**for** $k' = 1, 2, \cdots, K'$ **do**
    **for** $l = 1, 2, \cdots, L$ **do**
        Calculate $\mathbf{H}^{(l)}$ via Equation 2.3;
    **end for**
    $\mathbf{Z}^{k'} = \mathbf{H}^{(L)}$;
    **for** $l = 1, 2, \cdots, L$ **do**
        Calculate $\widehat{\mathbf{H}}^{(l)}$ via Equation 2.7;
    **end for**
    $\widehat{\mathbf{X}}^{k'} = \widehat{\mathbf{H}}^{(L)}$;
    Calculate $\widehat{\mathbf{A}}^{k'}$ via Equation 2.5;
    Calculate anomaly score $s_i^{k'}$ of each node $v_i^{k'}$ in $\mathcal{G}_{\text{test}}^{k'}$ via Equation 2.10.
**end for**

# Appendix B. Detailed Experimental Results of Phrase Embedding-Based Methods

The detailed results of the three judgment methods α, β, and γ are shown in Tables 4.1 - 4.3. In the tables, a phrase in parentheses represents that the explanation belongs to class 0. We discussed carefully in assigning a class label, e.g., both explanations on English score for men in type V were judged class 0, as we agree that it is widely known that men who are good at English exist as those who are not. FP and FN with bold fonts represent that the explanation is judged as a false positive and a false negative, respectively. A blank in the Result column either represents a true positive or

a true negative.

In method β, as we explained in Chapter 3.5.2, we specify $X'$ for each $X$. We denote the assignment in the form of $(X, X')$, though a phrase could be a set of phrases to save space. Refer to Table 4.2 for the following assignments. In type IV, ({Muslims, Judaisms}, Christians) and vice versa. Type V, (men, women) and vice versa. Type VI, ({Iranians, Afghans}, Americans) and ({Americans, French}, Iranians). Note that in this type, $X'$ is specified with the word "than" to one country. Type VII, (developing countries, advanced countries) and vice versa. Note that $X'$ is specified in the text. Type XII, (Asia, {Africa, Europe}) and vice versa. Type XIV, ({China, India, United States}, United Kingdom) and vice versa. Note that the United Kingdom is considered to be the representative of the countries with a small amount of CO2 emissions in the current era. Type XV, (large hospitals, small hospitals) and vice versa. Type XVI, ({Omicron strain, Alpha strain, Beta strain, Gamma strain}, Delta strain) and vice versa. Type XVII, (Africa, {Asia, Americas, Europe}), (Asia, {Africa, Americas, Europe}), (Americas, {Africa, Asia, Europe}), and (Europe, {Africa, Asia, Americas}). Note that this assignment results from using the expression "than other regions" in the explanation. Type XVIII, ({cancer, Alzheimer's disease, heart disease, pneumonia}, periodontal disease) and vice versa. Note that this assignment follows a similar reason to type XIV.

Note that in types II-1 and II-2, $X'$ is rather specified at the end of the explanation as "of the healthiest" or "of the unhealthiest", which has a fixed correspondence to $Y$, i.e., "poorest" or "richest", respectively. Thus for these types, we denote the assignment in the form of $(Y, X')$, which are (poorest, {Japan, Singapore}) and (richest, {Central African Republic, Somalia}). These 4 countries are selected as the representatives of the healthiest or unhealthiest as they are located in the upper-right or lower-left corners in Figure 3.2 II, respectively.

In method γ, as we explained in Chapter 3.5.3, we specify $Y'$ for each $Y$. The assignment is straightforward, as for each $Y$, $Y'$ consists of its variants belonging to the opposite class. For instance, as shown in Table 4.3, in type III, when $Y$ is "proportional to GDP", $Y'$ is {"inversely proportional to GDP", "not correlated to GDP"} and vice versa.

Table 4.1: Detailed results of method α.

| Type | $X$ | $Y$ | $\theta_{\text{relevance}}$ | $\theta_{\text{fear}}$ | $\theta_{\text{bad habit}}$ | Result |
|---|---|---|---|---|---|---|
| I | deep-fried food | pancreatic cancer | 2.309 | 11.5 | 2.070 | |
| 28-0 | | Alzheimer's disease | 1.968 | 25.3 | 2.070 | |
| | | periodontal disease | 1.715 | 7.10 | 2.070 | |
| | | (flu) | 0.941 | 0.00 | 2.070 | |
| | | (alopecia areata) | 2.736 | 0.60 | 2.070 | |
| | | (bone fracture) | 1.374 | 0.00 | 2.070 | |
| | | (nosebleeds) | 1.363 | 0.00 | 2.070 | |
| | alcohol abuse | pancreatic cancer | 1.921 | 11.5 | 1.755 | |
| | | Alzheimer's disease | 4.211 | 25.3 | 1.755 | |
| | | periodontal disease | 4.055 | 7.10 | 1.755 | |
| | | (flu) | 1.474 | 0.00 | 1.755 | |
| | | (alopecia areata) | 4.522 | 0.60 | 1.755 | |
| | | (bone fracture) | 1.852 | 0.00 | 1.755 | |
| | | (nosebleeds) | 1.933 | 0.00 | 1.755 | |
| | heavy drinking | pancreatic cancer | 1.624 | 11.5 | 1.528 | |
| | | Alzheimer's disease | 3.557 | 25.3 | 1.528 | |
| | | periodontal disease | 2.965 | 7.10 | 1.528 | |
| | | (flu) | 1.656 | 0.00 | 1.528 | |
| | | (alopecia areata) | 3.159 | 0.60 | 1.528 | |
| | | (bone fracture) | 1.893 | 0.00 | 1.528 | |
| | | (nosebleeds) | 1.962 | 0.00 | 1.528 | |
| | long distance running | (pancreatic cancer) | 0.048 | 11.5 | 0.922 | |
| | | (Alzheimer's disease) | 0.509 | 25.3 | 0.922 | |
| | | (periodontal disease) | -0.133 | 7.10 | 0.922 | |
| | | (flu) | 0.769 | 0.00 | 0.922 | |
| | | (alopecia areata) | 1.415 | 0.60 | 0.922 | |
| | | (bone fracture) | 1.509 | 0.00 | 0.922 | |
| | | (nosebleeds) | 1.262 | 0.00 | 0.922 | |

Table 4.2: Detailed results of method β, where the abbreviations MR, ER, and CO2E represent mortality rates, enrollment rates, and CO2 emissions, respectively.

| Type | $X$ | $Y$ | $\theta^{\beta}_{XY}$ | Result | Type | $X$ | $Y$ | $\theta^{\beta}_{XY}$ | Result |
|---|---|---|---|---|---|---|---|---|---|
| II-1 | Cuba | poorest | 0.273 | | XII | Asia | large amount of CO2E | 0.848 | |
| 5-3 | | (richest) | 0.266 | | 4-2 | | (small amount of CO2E) | 0.227 | |
| | Nicaragua | poorest | 0.245 | | | Africa | small amount of CO2E | 0.193 | **FN** |
| | | (richest) | 0.192 | | | | (large amount of CO2E) | 0.534 | |
| | Bangladesh | poorest | 0.267 | **FN** | | Europe | small amount of CO2E | 0.216 | **FN** |
| | | (richest) | 0.284 | | | | (large amount of CO2E) | 0.812 | |
| | North Korea | poorest | 0.258 | **FN** | XIV | China | large amount of CO2E | 0.951 | |
| | | (richest) | 0.314 | **FP** | 7-1 | | (small amount of CO2E) | 0.265 | |
| II-2 | United Arab Emirates | richest | 0.291 | **FN** | | India | large amount of CO2E | 0.741 | |
| 3-5 | | (poorest) | 0.185 | | | | (small amount of CO2E) | 0.435 | |
| | Qatar | richest | 0.244 | **FN** | | United States | large amount of CO2E | 0.520 | **FN** |
| | | (poorest) | 0.223 | | | | (small amount of CO2E) | 0.143 | |
| | Equatorial Guinea | richest | 0.238 | **FN** | | United Kingdom | (small amount of CO2E) | 0.481 | |
| | | (poorest) | 0.180 | | | | (large amount of CO2E) | 0.570 | |
| | Botswana | richest | 0.271 | **FN** | XV | small hospitals | dangerous hospitals | 0.826 | **FN** |
| | | (poorest) | 0.278 | **FP** | 0-4 | | (safe hospitals) | 0.861 | **FP** |
| IV | Muslims | many babies | 0.536 | **FN** | | large hospitals | safe hospitals | 0.860 | **FN** |
| 3-3 | | (few babies) | 0.561 | **FP** | | | (dangerous hospitals) | 0.867 | **FP** |
| | Judaisms | many babies | 0.621 | | XVI | Omicron strain | less dangerous | 1.012 | **FN** |
| | | (few babies) | 0.570 | | 1-9 | | (more dangerous) | 1.138 | **FP** |
| | Christians | few babies | 0.552 | **FN** | | Alpha strain | less dangerous | 0.855 | **FN** |
| | | (many babies) | 0.526 | | | | (more dangerous) | 0.935 | **FP** |
| V | women | low math score | 0.020 | **FN** | | Beta strain | less dangerous | 1.023 | **FN** |
| 7-1 | | (high math score) | 0.299 | | | | (more dangerous) | 1.155 | **FP** |
| | men | high math score | 0.327 | | | Gamma strain | less dangerous | 0.812 | **FN** |
| | | (low math score) | 0.163 | | | | (more dangerous) | 0.845 | **FP** |
| | women | (low English score) | 0.106 | | | Delta strain | more dangerous | 0.672 | **FN** |
| | | high English score | 0.270 | | | | (less dangerous) | 0.709 | |
| | men | (high English score) | 0.199 | | XVII | Africa | low GDP | 0.809 | **FN** |
| | | (low English score) | 0.081 | | 5-3 | | (high GDP) | 0.875 | |
| VI | Iranians | many children | 0.597 | **FN** | | Asia | high GDP | 1.009 | |
| 5-3 | | (few children) | 0.616 | **FP** | | | (low GDP) | 0.709 | |
| | Afghans | many children | 0.634 | | | Americas | high GDP | 0.884 | **FN** |
| | | (few children) | 0.593 | | | | (low GDP) | 0.699 | |
| | French | few children | 0.670 | | | Europe | high GDP | 0.955 | **FN** |
| | | (many children) | 0.662 | | | | (low GDP) | 0.776 | |
| | Americans | few children | 0.505 | **FN** | XVIII | cancer | (long life expectancy) | 0.910 | |
| | | (many children) | 0.531 | | 4-6 | | short life expectancy | 1.027 | **FN** |
| VII | developing countries | high infant MR | 1.264 | **FN** | | Alzheimer's disease | (long life expectancy) | 0.972 | |
| 3-5 | | (low infant MR) | 1.365 | **FP** | | | short life expectancy | 1.108 | **FN** |
| | advanced countries | low infant MR | 1.355 | **FN** | | heart disease | (long life expectancy) | 0.901 | |
| | | (high infant MR) | 1.407 | **FP** | | | short life expectancy | 1.066 | **FN** |
| | developing countries | low ER | 1.206 | **FN** | | pneumonia | (long life expectancy) | 0.926 | |
| | | (high ER) | 1.272 | | | | short life expectancy | 1.107 | **FN** |
| | advanced countries | high ER | 1.484 | | | periodontal disease | (short life expectancy) | 1.500 | **FP** |
| | | (low ER) | 1.033 | | | | long life expectancy | 1.224 | **FN** |

Table 4.3: Detailed results of method γ, where the abbreviations ED, IA, and ND, represent epidemic damages, industrial accidents, and natural disasters, respectively.

| Type | $X$ | $Y$ | $\theta^{\gamma}_{XY}$ | Result |
|---|---|---|---|---|
| III | life expectancy | proportional to GDP | 0.144 | |
| 6-0 | | (inversely proportional to GDP) | 0.033 | |
| | | (not correlated to GDP) | 0.079 | |
| | healthy life | proportional to GDP | 0.168 | |
| | expectancy | (inversely proportional to GDP) | 0.059 | |
| | | (not correlated to GDP) | 0.150 | |
| VIII | child labor | not decreasing | 0.061 | **FN** |
| 9-6 | | (decreasing) | 0.062 | |
| | | (not increasing) | 0.090 | |
| | | (increasing) | 0.111 | **FP** |
| | | (constant) | 0.111 | |
| | child hunger | not decreasing | 0.146 | **FN** |
| | | (decreasing) | 0.148 | |
| | | (not increasing) | 0.187 | |
| | | (increasing) | 0.177 | |
| | | (constant) | 0.188 | **FP** |
| | child mortality | not decreasing | 0.195 | **FN** |
| | | (decreasing) | 0.206 | |
| | | (not increasing) | 0.207 | |
| | | (increasing) | 0.217 | **FP** |
| | | (constant) | 0.127 | |
| IX | world population | will just increase | 0.203 | **FN** |
| 4-1 | | will rapidly increase | 0.221 | |
| | | (will just decrease) | 0.130 | |
| | | (will rapidly decrease) | 0.126 | |
| | | (will keep constant) | 0.211 | |
| X | increasing | increasing in deaths from ND | 0.898 | |
| 27-0 | in ND | (decreasing in deaths from ND) | 0.820 | |
| | | (constant deaths from ND) | 0.650 | |
| | decreasing | (increasing in deaths from ND) | 0.793 | |
| | in ND | (decreasing in deaths from ND) | 0.896 | |
| | | (constant deaths from ND) | 0.638 | |
| | constant | (increasing in deaths from ND) | 0.636 | |
| | ND | (decreasing in deaths from ND) | 0.563 | |
| | | (constant deaths from ND) | 0.844 | |
| | increasing | increasing in deaths from ED | 0.940 | |
| | in ED | (decreasing in deaths from ED) | 0.851 | |
| | | (constant deaths from ED) | 0.801 | |
| | decreasing | (increasing in deaths from ED) | 0.881 | |
| | in ED | (decreasing in deaths from ED) | 0.934 | |
| | | (constant deaths from ED) | 0.787 | |
| X | constant ED | (increasing in deaths from ED) | 0.838 | |
| | | (decreasing in deaths from ED) | 0.778 | |
| | | (constant deaths from ED) | 0.915 | |
| | increasing in IA | increasing in deaths from IA | 0.945 | |
| | | (decreasing in deaths from IA) | 0.891 | |
| | | (constant deaths from IA) | 0.791 | |
| | decreasing in IA | (increasing in deaths from IA) | 0.862 | |
| | | (decreasing in deaths from IA) | 0.942 | |
| | | (constant deaths from IA) | 0.752 | |
| | constant IA | (increasing in deaths from IA) | 0.764 | |
| | | (decreasing in deaths from IA) | 0.714 | |
| | | (constant deaths from IA) | 0.879 | |
| XI | death of many babies | increasing | 0.133 | |
| 20-0 | | (decreasing) | 0.114 | |
| | | (not increasing) | 0.129 | |
| | | (not decreasing) | 0.112 | |
| | | (constant) | 0.064 | |
| | death of many children | increasing | 0.129 | |
| | | (decreasing) | 0.102 | |
| | | (not increasing) | 0.116 | |
| | | (not decreasing) | 0.092 | |
| | | (constant) | 0.076 | |
| | death of many adults | increasing | 0.169 | |
| | | (decreasing) | 0.114 | |
| | | (not increasing) | 0.158 | |
| | | (not decreasing) | 0.141 | |
| | | (constant) | 0.052 | |
| | death of many old | increasing | 0.156 | |
| | people | (decreasing) | 0.122 | |
| | | (not increasing) | 0.138 | |
| | | (not decreasing) | 0.130 | |
| | | (constant) | 0.030 | |
| XIII | risk of death from | increasing | 0.082 | |
| 9-0 | cancer | (decreasing) | 0.033 | |
| | | (constant) | -0.012 | |
| | risk of death from | increasing | 0.064 | |
| | Alzheimer's disease | (decreasing) | 0.018 | |
| | | (constant) | -0.042 | |
| | risk of death from | increasing | 0.094 | |
| | heart disease | (decreasing) | 0.063 | |
| | | (constant) | 0.006 | |

# Bibliography


[1] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully Convolutional Localization Networks for Dense Captioning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4565–4574, 2016.

[2] H. Rosling, O. Rosling, and A. R. Roennlund, *Factfulness: Ten Reasons We're Wrong about the World - and Why Things are Better than You Think*. London: Sceptre, 2018.

[3] M. Blastland and D. Spiegelhalter, *The Norm Chronicles: Stories and Numbers About Danger and Death*. New York: Basic Books, 2014.

[4] O. Rosling, A. R. Rönnlund, and H. Rosling, "Gapminder Download the Data." `https://www.gapminder.org/data/`, 2005.

[5] A. Arning, R. Agrawal, and P. Raghavan, "A Linear Method for Deviation Detection in Large Databases," in *Proc. International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 164–169, 1996.

[6] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "A Survey of Single-Scene Video Anomaly Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44, No. 5, pp. 2293–2312, 2020.

[7] M. F. Fadjrimiratno, Y. Hatae, T. Matsukawa, and E. Suzuki, "Detecting Anomalies from Human Activities by an Autonomous Mobile Robot Based on "Fast and Slow" Thinking," in *Proc. International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), Vol. 5: VISAPP*, pp. 943–953, 2021.

[8] O. P. Popoola and K. Wang, "Video-Based Abnormal Human Behavior Recognition—A Review," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 42, No. 6, pp. 865–878, 2012.

[9] C. Esterwood and L. P. Robert, "Personality in Healthcare Human Robot Interaction (H-HRI) a Literature Review and Brief Critique," in *Proc. International Conference on Human-Agent Interaction (HAI)*, pp. 87–95, 2020.

[10] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6479–6488, 2018.

[11] M. Arvan, "Mental Time-Travel, Semantic Flexibility, and AI Ethics," *AI & SOCIETY*, Vol. 1, pp. 1–20, 2018.

[12] T. S. Doherty and A. E. Carroll, "Believing in Overcoming Cognitive Biases," *AMA Journal of Ethics*, Vol. 22, No. 9, pp. 773–778, 2020.

[13] T. Neal, P. Lienert, E. Denne, and J. P. Singh, "A General Model of Cognitive Bias in Human Judgment and Systematic Review Specific to Forensic Mental Health," *Law and Human Behavior*, Vol. 46, No. 2, pp. 99–120, 2022.

[14] N. Dong and E. Suzuki, "GIAD: Generative Inpainting-Based Anomaly Detection via Self-Supervised Learning for Human Monitoring," in *Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, pp. 418–432, 2021.

[15] A. M. Braşoveanu and R. Andonie, "Integrating Machine Learning Techniques in Semantic Fake News Detection," *Neural Processing Letters*, Vol. 53, No. 5, pp. 3055–3072, 2021.

[16] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media," *Big Data*, Vol. 8, No. 3, pp. 171–188, 2020.

[17] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2931–2937, 2017.

[18] K. Zhang, H. Shinden, T. Mutsuro, and E. Suzuki, "Judging Instinct Exploitation in Statistical Data Explanations Based on Word Embedding," in *Proc. AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 867–879, 2022.

[19] Chen, Peter Pin-Shan, "The Entity-Relationship Model—Toward a Unified View of Data," *ACM Transactions on Database Systems*, Vol. 1, No. 1, pp. 9–36, 1976.

[20] P. Lin, S. Yu, X. Zhou, P. Peng, K. Li, and X. Liao, "Community Search over Large Semantic-Based Attribute Graphs," *World Wide Web*, Vol. 25, No. 2, pp. 927–948, 2022.

[21] Y. Hatae, Q. Yang, M. F. Fadjrimiratno, Y. Li, T. Matsukawa, and E. Suzuki, "Detecting Anomalous Regions from an Image Based on Deep Captioning," in *Proc. International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), Vol. 5: VISAPP*, pp. 326–335, 2020.

[22] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," in *Proc. International Conference on Information Processing in Medical Imaging (IPMI)*, pp. 146–157, 2017.

[23] P. Seeböck, S. Waldstein, S. Klimscha, B. S. Gerendas, R. Donner, T. Schlegl, U. Schmidt-Erfurth, and G. Langs, "Identifying and Categorizing Anomalies in Retinal Imaging Data," *arXiv preprint arXiv:1612.00686*, 2016.

[24] M. J. Choi, A. Torralba, and A. S. Willsky, "Context Models and Out-of-Context Objects," *Pattern Recognition Letters*, Vol. 33, No. 7, pp. 853–862, 2012.

[25] A. Pasini and E. Baralis, "Detecting Anomalies in Image Classification by Means of Semantic Relationships," in *Proc. IEEE International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pp. 231–238, 2019.

[26] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How Powerful Are Graph Neural Networks?," *arXiv preprint arXiv:1810.00826*, 2018.

[27] A. L. Casebeer and M. J. Verhoef, "Combining Qualitative and Quantitative Research Methods: Considering the Possibilities for Enhancing the Study of Chronic Diseases," *Chronic Diseases in Canada*, Vol. 18, No. 3, pp. 130–135, 1997.

[28] C. Williams *et al.*, "Research Methods," *Journal of Business & Economics Research*, Vol. 5, No. 3, 2007.

[29] J. Noyes, A. Booth, G. Moore, K. Flemming, Ö. Tunçalp, and E. Shakibazadeh, "Synthesising Quantitative and Qualitative Evidence to Inform Guidelines on Complex Interventions: Clarifying the Purposes, Designs and Outlining Some Methods," *BMJ Global Health*, Vol. 4, No. Suppl 1, Article e000893, 2019.

[30] R. T. Collins, A. J. Lipton, and T. Kanade, "Introduction to the Special Section on Video Surveillance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 745–746, 2000.

[31] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, "Human Behavior Analysis in Video Surveillance: A Social Signal Processing Perspective," *Neurocomputing*, Vol. 100, pp. 86–97, 2013.

[32] S. C. Mukhopadhyay, "Wearable Sensors for Human Activity Monitoring: A Review," *IEEE Sensors Journal*, Vol. 15, No. 3, pp. 1321–1330, 2014.

[33] G. Liu, Q. Zhang, Y. Cao, G. Tian, and Z. Ji, "Online Human Action Recognition with Spatial and Temporal Skeleton Features Using a Distributed Camera Network," *International Journal of Intelligent Systems*, Vol. 36, No. 12, pp. 7389–7411, 2021.

[34] M. Kashef, A. Visvizi, and O. Troisi, "Smart City as a Smart Service System: Human-Computer Interaction and Smart City Surveillance Systems," *Computers in Human Behavior*, Vol. 124, Article 106923, 2021.

[35] W. Luo, W. Liu, D. Lian, and S. Gao, "Future Frame Prediction Network for Video Anomaly Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44, No. 11, pp. 7505–7520, 2021.

[36] M. Yu, G. Li, D. Jiang, G. Jiang, B. Tao, and D. Chen, "Hand Medical Monitoring System Based on Machine Learning and Optimal EMG Feature Set," *Personal and Ubiquitous Computing*, Vol. 1, pp. 1–17, 2019.

[37] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 322–339, 2020.

[38] Y. Deguchi, D. Takayama, S. Takano, V.-M. Scuturici, J.-M. Petit, and E. Suzuki, "Skeleton Clustering by Multi-Robot Monitoring for Fall Risk Discovery," *Journal of Intelligent Information Systems*, Vol. 48, pp. 75–115, 2017.

[39] F. Meng, G. Yuan, S. Lv, Z. Wang, and S. Xia, "An Overview on Trajectory Outlier Detection," *Artificial Intelligence Review*, Vol. 52, pp. 2437–2456, 2019.

[40] N. Dong and E. Suzuki, "GIAD-ST: Detecting Anomalies in Human Monitoring Based on Generative Inpainting via Self-Supervised Multi-Task Learning," *Journal of Intelligent Information Systems*, pp. 1–22, 2022.

[41] J. Yi and S. Yoon, "Patch SVDD: Patch-Level SVDD for Anomaly Detection and Segmentation," in *Proc. Asian Conference on Computer Vision (ACCV)*, 2020.

[42] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-Supervised Learning for Anomaly Detection and Localization," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9664–9674, 2021.

[43] K. Wu, L. Zhu, W. Shi, W. Wang, and J. Wu, "Self-Attention Memory-Augmented Wavelet-CNN for Anomaly Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 33, No. 3, pp. 1374–1385, 2023.

[44] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure Inference Net: Object Detection Using Scene-Level Context and Instance-Level Relationships," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6985–6994, 2018.

[45] X. Wang and A. Gupta, "Videos as Space-Time Region Graphs," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 399–417, 2018.

[46] C. Sun, Y. Jia, Y. Hu, and Y. Wu, "Scene-Aware Context Reasoning for Unsupervised Abnormal Event Detection in Videos," in *Proc. ACM International Conference on Multimedia (MM)*, pp. 184–192, 2020.

[47] H. Mu, R. Sun, M. Wang, and Z. Chen, "Spatio-Temporal Graph-Based CNNs for Anomaly Detection in Weakly-Labeled Videos," *Information Processing & Management*, Vol. 59, No. 4, p. 102983, 2022.

[48] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *arXiv preprint arXiv:1609.02907*, 2016.

[49] K. Ding, J. Li, R. Bhanushali, and H. Liu, "Deep Anomaly Detection on Attributed Networks," in *Proc. SIAM International Conference on Data Mining (SDM)*, pp. 594–602, 2019.

[50] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD–A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9592–9600, 2019.

[51] S. Venkataramanan, K.-C. Peng, R. V. Singh, and A. Mahalanobis, "Attention Guided Anomaly Localization in Images," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 485–503, 2020.

[52] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, "Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection," in *Proc. International Conference on Computer Vision (ICCV)*, pp. 1705–1714, 2019.

[53] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 32, No. 1, pp. 4–24, 2020.

[54] Y. Liu, Z. Li, S. Pan, C. Gong, C. Zhou, and G. Karypis, "Anomaly Detection on Attributed Networks via Contrastive Self-Supervised Learning," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 33, No. 6, pp. 2378–2392, 2021.

[55] Y. Zheng, M. Jin, Y. Liu, L. Chi, K. T. Phan, and Y.-P. P. Chen, "Generative and Contrastive Self-Supervised Learning for Graph Anomaly Detection," *IEEE Transactions on Knowledge and Data Engineering*, 2021, in press.

[56] M. Jin, Y. Liu, Y. Zheng, L. Chi, Y.-F. Li, and S. Pan, "Anemone: Graph Anomaly Detection with Multi-Scale Contrastive Learning," in *Proc. In-*

*ternational Conference on Information & Knowledge Management (CIKM)*, pp. 3122–3126, 2021.

[57] H. Fan, F. Zhang, and Z. Li, "Anomalydae: Dual Autoencoder for Anomaly Detection on Attributed Networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5685–5689, 2020.

[58] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph Attention Networks," *arXiv preprint arXiv:1710.10903*, 2017.

[59] A. Salehi and H. Davulcu, "Graph Attention Auto-Encoders," *arXiv preprint arXiv:1905.10715*, 2019.

[60] M. Z. Zaheer, J.-h. Lee, M. Astrid, and S.-I. Lee, "Old Is Gold: Redefining the Adversarially Learned One-Class Classifier Training Paradigm," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14183–14193, 2020.

[61] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training," in *Proc. Asian Conference on Computer Vision (ACCV)*, pp. 622–637, 2019.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[63] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *arXiv preprint arXiv:1908.10084*, 2019.

[64] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[65] Z. Liu, W. Lin, Y. Shi, and J. Zhao, "A Robustly Optimized BERT Pre-training Approach with Post-training," in *Proc. Chinese National Conference on Computational Linguistics (CCL)*, pp. 471–484, 2021.

[66] T. N. Kipf and M. Welling, "Variational Graph Auto-Encoders," *arXiv preprint arXiv:1611.07308*, 2016.

[67] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proc. International Conference on Machine Learning (ICML)*, pp. 807–814, 2010.

[68] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *International Journal of Computer Vision*, Vol. 123, No. 1, pp. 32–73, 2017.

[69] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

[70] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-Wise Training of Deep Networks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 153–160, 2007.

[71] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[72] J. Shlens, "A Tutorial on Principal Component Analysis," *arXiv preprint arXiv:1404.1100*, 2014.

[73] M. Westerlund, "The Emergence of Deepfake Technology: A Review," *Technology Innovation Management Review*, Vol. 9, No. 11, pp. 40–53, 2019.

[74] T. Zemčík, "Failure of Chatbot Tay Was Evil, Ugliness and Uselessness in Its Nature or Do We Judge It Through Cognitive Shortcuts and Biases?," *AI & SOCIETY*, Vol. 36, pp. 361–367, 2021.

[75] H. H. Thorp, "ChatGPT Is Fun, but Not an Author," *Science*, Vol. 379, No. 6630, pp. 313–313, 2023.

[76] M. Liebrenz, R. Schleifer, A. Buadze, D. Bhugra, and A. Smith, "Generating Scholarly Content with ChatGPT: Ethical Challenges for Medical Publishing," *The Lancet Digital Health*, Vol. 5, No. 3, pp. e105–e106, 2023.

[77] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, "Exploring AI Ethics of ChatGPT: A Diagnostic Analysis," *arXiv preprint arXiv:2301.12867*, 2023.

[78] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013.

[79] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global Vectors for Word Representation," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[80] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep Contextualized Word Representations," *arXiv preprint arXiv:1802.05365*, 2018.

[81] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language Models Are Unsupervised Multitask Learners," *OpenAI Blog*, Vol. 1, No. 8, p. 9, 2019.

[82] D. A. Scheufele and N. M. Krause, "Science Audiences, Misinformation, and Fake News," *Proceedings of the National Academy of Sciences*, Vol. 116, No. 16, pp. 7662–7669, 2019.

[83] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," *ACM SIGKDD Explorations Newsletter*, Vol. 19, No. 1, pp. 22–36, 2017.

[84] J. C. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Supervised Learning for Fake News Detection," *IEEE Intelligent Systems*, Vol. 34, No. 2, pp. 76–81, 2019.

[85] A. Vlachos and S. Riedel, "Fact Checking: Task Definition and Dataset Construction," in *Proc. ACL Workshop on Language Technologies and Computational Social Ccience*, pp. 18–22, 2014.

[86] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect Rumors Using Time Series of Social Context Information on Microblogging Websites," in *Proc. ACM on Conference on Information and Knowledge Management (CIKM)*, pp. 1751–1754, 2015.

[87] L. Wu and H. Liu, "Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate," in *Proc. International Conference on Web Search and Data Mining (WSDM)*, pp. 637–645, 2018.

[88] W. Y. Wang, ""Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection," *arXiv preprint arXiv:1705.00648*, 2017.

[89] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li, "Exploring the Role of Visual Content in Fake News Detection," *arXiv preprint arXiv:2003.05096*, 2020.

[90] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal Variational Autoencoder for Fake News Detection," in *Proc. The World Wide Web Conference (WWW)*, pp. 2915–2921, 2019.

[91] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection," in *Proc. International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 849–857, 2018.

[92] L. M. Lesser and E. Nordenhaug, "Ethical Statistics and Statistical Ethics: Making an Interdisciplinary Module," *Journal of Statistics Education*, Vol. 12, No. 3, 2004.

[93] G. Tripepi, K. J. Jager, F. W. Dekker, and C. Zoccali, "Selection Bias and Information Bias in Clinical Research," *Nephron Clinical Practice*, Vol. 115, No. 2, pp. c94–c99, 2010.

[94] M. S. Thiese, Z. C. Arnold, and S. D. Walker, "The Misuse and Abuse of Statistics in Biomedical Research," *Biochemia Medica*, Vol. 25, No. 1, pp. 5–11, 2015.

[95] M. É. Czeisler, J. F. Wiley, C. A. Czeisler, S. M. Rajaratnam, and M. E. Howard, "Uncovering Survivorship Bias in Longitudinal Mental Health Surveys During

the COVID-19 Pandemic," *Epidemiology and Psychiatric Sciences*, Vol. 30, Article e45, 2021.

[96] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning, "Bilingual Word Embeddings for Phrase-Based Machine Translation," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1393–1398, 2013.

[97] S. Kim, N. Fiorini, W. J. Wilbur, and Z. Lu, "Bridging the Gap: Incorporating a Semantic Similarity Measure for Effectively Mapping Pubmed Queries to Documents," *Journal of Biomedical Informatics*, Vol. 75, pp. 122–127, 2017.

[98] I. Lopez-Gazpio, M. Maritxalar, A. Gonzalez-Agirre, G. Rigau, L. Uria, and E. Agirre, "Interpretable Semantic Textual Similarity: Finding and Explaining Differences Between Sentences," *Knowledge-Based Systems*, Vol. 119, pp. 186–199, 2017.

[99] Y. Li, Z. A. Bandar, and D. McLean, "An Approach for Measuring Semantic Similarity Between Words Using Multiple Information Sources," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 4, pp. 871–882, 2003.

[100] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring Semantic Similarity Between Words Using Web Search Engines," in *Proc. International Conference on World Wide Web (WWW)*, pp. 757–766, 2007.

[101] R. L. Cilibrasi and P. M. Vitanyi, "The google similarity distance," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, No. 3, pp. 370–383, 2007.

[102] D. Chandrasekaran and V. Mago, "Evolution of Semantic Similarity—a Survey," *ACM Computing Surveys*, Vol. 54, No. 2, pp. 1–37, 2021.

[103] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *Proc. International Conference on Machine Learning (ICML)*, pp. 1188–1196, 2014.

[104] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics Derived Automatically From Language Corpora Contain Human-Like Biases," *Science*, Vol. 356, No. 6334, pp. 183–186, 2017.

[105] W. Guo and A. Caliskan, "Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-Like Biases," in *Proc. AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 122–133, 2021.

[106] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised Learning of Universal Sentence Representations From Natural Language Inference Data," *arXiv preprint arXiv:1705.02364*, 2017.

[107] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, *et al.*, "Universal Sentence Encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[108] K. Wu, S. Yang, and K. Q. Zhu, "False Rumors Detection on Sina Weibo by Propagation Structures," in *Proc. International Conference on Data Engineering (ICDE)*, pp. 651–662, 2015.

[109] B. Balcerzak, W. Jaworski, and A. Wierzbicki, "Application of TextRank Algorithm for Credibility Assessment," in *Proc. IEEE/WIC International Joint*

*Conferences on Web Intelligence and Intelligent Agent Technologies (WI-IAT)*, pp. 451–454, 2014.

[110] A. Kazemi, V. Pérez-Rosas, and R. Mihalcea, "Biased TextRank: Unsupervised Graph-Based Content Extraction," in *Proc. International Conference on Computational Linguistics (COLING)*, pp. 1642–1652, 2020.

[111] Q. Mao, Y. Wang, C. Yang, L. Du, H. Peng, J. Wu, J. Li, and Z. Wang, "HiGIL: Hierarchical Graph Inference Learning for Fact Checking," in *Proc. International Conference on Data Mining (ICDM)*, pp. 329–337, 2022.

[112] R. Mihalcea and P. Tarau, "Textrank: Bringing Order into Text," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 404–411, 2004.

[113] J. Kim and K.-s. Choi, "Unsupervised Fact Checking by Counter-Weighted Positive and Negative Evidential Paths in A Knowledge Graph," in *Proc. International Conference on Computational Linguistics (COLING)*, pp. 1677–1686, 2020.

[114] N. Vedula and S. Parthasarathy, "FACE-KEG: FAct Checking Explained using KnowledgE Graphs," in *Proc. International Conference on Web Search and Data Mining (WSDM)*, pp. 526–534, 2021.

[115] L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang, and D. Lee, "DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation," in *Proc. International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 492–502, 2020.

[116] G. Luo, J. Li, J. Su, H. Peng, C. Yang, L. Sun, P. S. Yu, and L. He, "Graph Entropy Guided Node Embedding Dimension Selection for Graph Neural Networks," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2767–2774, 2021.

[117] M. Dehmer, "Information Processing in Complex Networks: Graph Entropy and Information Functionals," *Applied Mathematics and Computation*, Vol. 201, No. 1-2, pp. 82–94, 2008.

[118] C. C. Noble and D. J. Cook, "Graph-Based Anomaly Detection," in *Proc. International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 631–636, 2003.

[119] B. Sen, S.-H. Chu, and K. K. Parhi, "Ranking Regions, Edges and Classifying Tasks in Functional Brain Graphs by Sub-Graph Entropy," *Scientific Reports*, Vol. 9, No. 1, Article 7628, 2019.

[120] O. Araque, G. Zhu, and C. A. Iglesias, "A Semantic Similarity-Based Perspective of Affect Lexicons for Sentiment Analysis," *Knowledge-Based Systems*, Vol. 165, pp. 346–359, 2019.

[121] W. Chen, L. Chen, Y. Xie, W. Cao, Y. Gao, and X. Feng, "Multi-Range Attentive Bicomponent Graph Convolutional Network for Traffic Forecasting," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3529–3536, 2020.

[122] J. Deng, Y. Deng, and K. H. Cheong, "Combining Conflicting Evidence Based on Pearson Correlation Coefficient and Weighted Graph," *International Journal of Intelligent Systems*, Vol. 36, No. 12, pp. 7443–7460, 2021.

[123] H. Toivonen, F. Zhou, A. Hartikainen, and A. Hinkka, "Compression of Weighted Graphs," in *Proc. International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 965–973, 2011.

# Published Papers

1 **Kang Zhang**, Muhammad Fikko Fadjrimiratno, and Einoshin Suzuki. Context-Based Anomaly Detection via Spatial Attributed Graphs in Human Monitoring. in *Proc. International Conference on Neural Information Processing (ICONIP)*, pp. 450–463, 2021.

2 **Kang Zhang**, Hiroaki Shinden, Tatsuki Mutsuro, and Einoshin Suzuki. Judging Instinct Exploitation in Statistical Data Explanations Based on Word Embedding. in *Proc. AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 867–879, 2022.

3 **Kang Zhang**, Muhammad Fikko Fadjrimiratno, and Einoshin Suzuki. Region Anomaly Detection via Spatial and Semantic Attributed Graph in Human Monitoring. *Sensors*, Vol. 23, No. 3, Article 1307, 2023.

4 **Kang Zhang** and Einoshin Suzuki. Judging Credible and Unethical Statistical Data Explanations via Phrase Similarity Graph. in *Proc. Pacific Asia Conference on Information Systems (PACIS)*, paper 121, 2023.