https://doi.org/10.1016/j.apenergy.2019.113395

A Graph Mining-based Methodology for Discovering and Visualizing High-level Knowledge for Building Energy Management

Cheng Fan[1,2], Fu Xiao[2,*], Mengjie Song[3], Jiayuan Wang[1]

[1]Sino-Australia Joint Research Center in BIM and Smart Construction, Shenzhen University, Shenzhen, China

[2]Department of Building Services Engineering, The Hong Kong Polytechnic University, Hong Kong, China

[3]Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, Japan

[*]E-mail: linda.xiao@polyu.edu.hk; Tel.: (852) 27664194

Abstract

Building operations have evolved to be not only energy-intensive, but also information-intensive. Advanced data-driven methodologies are urgently needed to facilitate the tasks in building energy management. Currently, there are two main bottlenecks in analyzing building operational data. Firstly, few methodologies are available to represent and analyze data with complicated structures. Conventional data analytics are capable of analyzing information stored in a single two-dimensional data table, while lacking the ability to handle multi-relational databases. Secondly, it is still challenging to visualize the analysis results in a generic and flexible fashion, making it ineffective for knowledge interpretations and applications. As a promising solution, graphs can integrate and represent various types of information, providing promising approaches for the knowledge discovery from massive building operational data. This study proposes a novel graph-based methodology to analyze building operational data. The methodology consists of various stages and provides solutions for data exploration, graph generations, knowledge discovery and post-mining. It has been applied to analyze the actual building operational data of a public building in Hong Kong. The research results validate the potential of the graph-based methodology in characterizing high-level building operation patterns and atypical operations.

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

Keywords: Building operational data analysis; Unsupervised data mining; Graph mining; Frequent subgraph mining; Anomaly detection.

1. Introduction

Building operation has two prominent characteristics, i.e., energy-intensive and information-intensive. On the one hand, it accounts for 80-90% of the total energy consumption across the building life cycle, out of which 20-30% can be saved by applying advanced building automation technologies [1, 2]. On the other hand, the wide adoption of building automation systems has collected massive amounts of building operational data, making it feasible to develop data-driven approaches for building energy management. To fully embrace the power of information technologies, advanced data-driven methodologies are urgently needed to facilitate the decision-making during building operations.

To enhance the efficiency and effectiveness in utilizing massive amounts of building operational data, researchers have adopted various data mining techniques as the analysis tools. Relevant studies can be broadly classified into two types, i.e., supervised and unsupervised data mining-based studies. The first adopts supervised data mining to discover predictive knowledge, which describe the underlying relationships between input and output variables, are used as knowledge [3, 4]. The output variables were typically set as the building cooling or heating loads [5], electricity power consumptions [6], system performance indices [7, 8] and indoor environments [9, 10]. Existing studies in the building field mainly applied single model-based methods for model development [11]. The most widely used supervised learning techniques include artificial neural networks [12] and support vector machines [13]. Ensemble learning, which generates predictions based on a set of base models, has been utilized to further enhance the prediction performance [14]. The tree-based ensemble learning techniques, such as gradient boosting trees and random forests, have proved to be very useful for improving prediction accuracies [15]. Considering that building operational data are in essence time series data with high complexity, emerging data analytics, such as deep

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

autoencoders [16], deep recurrent networks [17] and generative adversarial networks [16, 18], have been used to optimize the overall process of predictive modelling. The predictive knowledge discovered is represented as predictive models, which serve as the basis for model-based fault detection and control optimization [19, 20]. The second adopts unsupervised data mining techniques to discover descriptive knowledge from building operational data [21, 22]. Clustering analysis has been widely applied to identify the intrinsic data structures or clusters [23, 24]. It has been typically used as a data pre-processing tool, as the clustering results are helpful to divide the operational data into different groups for separate analyses [25]. Association rule mining has gained its popularity in discovering associations among building variables [26, 27]. Conventional association rule mining algorithms, such as Apriori and FP-growth, are only capable of discovering static associations among categorical variables [28]. To overcome these limitations, researchers have explored the use of quantitative association rule mining [29], sequential pattern mining [30, 31] and gradual rule mining [32] in analyzing building operational data. The descriptive knowledge discovered can be used to describe the dynamic associations among building variables and has been successfully used to develop energy conservation measures [21, 22].

Despite encouraging results obtained, there are two bottlenecks in analyzing massive building operational data. Firstly, the above-mentioned data mining techniques are only compatible with building information stored in a single two-dimensional data table. Such data format typically uses columns to represent variables and rows to represent time steps. In such a case, the knowledge discovered cannot represent multi-level information, e.g., the spatial correlation and hierarchical structures among different building services components. Secondly, it is still challenging to visualize the data analysis results in a generic and flexible fashion. For the ease of human interpretation, specific post-mining and visualization methods are urgently needed to transform the raw data analysis results into interpretable insights. It can be foreseen that building operational data will become more diverse and complex due to the enrichment in the

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

types of information that can be collected, e.g., temporal and spatial information. Advanced data analytics are therefore needed to ensure the mining efficiency and effectiveness, while providing generic and flexible solutions for insight interpretations and applications.

To tackle the above-mentioned bottlenecks, this study proposes a novel graph-based methodology to analyze building operational data. Graphs can represent a broad spectrum of information which takes various formats ranging from traditional vectors to time-series, spatial information, hierarchical structures and etc. It has the ability to integrate and represent complicated interactions among building variables and therefore, providing a promising approach for insight extractions. This study attempts to investigate the potential of graph data and graph mining techniques in the building field. A graph-based methodology is proposed to analyze building operational data. It consists of various stages and provides solutions for data exploration, graph generations, knowledge discovery and post-mining. The paper is organized as follows. Section 2 provides the theoretical background. Section 3 describes the research methodology. The case study is shown in Section 4 and conclusions are drawn in Section 5.

2. Theoretical background

2.1 Basics on graph data

Graph is one of the most generic, natural and interpretable formats for information representations. It is regarded as one of the most widely used formats for representing complicated and multi-relational data [33]. A graph G consists of a set of vertices (or nodes) denoted as $V(G)$ and a set of edges (or links) denoted as $E(G)$. A graph S is a subgraph of graph G if $V(S) \subseteq V(G)$ and $E(S) \subseteq E(G)$. A vertex usually represents an entity, while the edges describe the relationships among vertices. Graphs can be either directed or undirected, depending on whether the edges have directions or not. Graphs can provide great flexibility for knowledge discovery as users can readily design and manipulate graph layouts for integrating and representing various types of information, ranging from vector data to time series, spatial information and hierarchical structures.

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

An example is given to illustrate the potential of graphs for information integrations and representations. Table-1 presents the power consumptions of a chiller and a cooling tower at time T_1 and T_2. Table-2 records the spatial location of these two components, i.e., one in basement and the other on rooftop. It is not easy to integrate these two tables into one without information loss, as there are two types of information, e.g., temporal and spatial information. By contrast, a graph can be readily designed for information integrations as shown in Fig. 1. The top 2 vertices represent the temporal information and are labelled as "T_1" and "T_2" respectively. The edge connecting these two vertices are labelled as "*dT=1*", which indicates the temporal difference. Each of the top 2 vertices is connected with two vertices labelled as "*Chiller*" and "*CT*". The power consumptions are encoded as edge labels. The bottom two vertices stand for the spatial information. It should be noted that graphs are highly flexible and different graph layouts can be used to represent the same piece of information.

Table-1 The chiller and cooling tower power consumptions

Time/Power	Chiller	Cooling tower
T1	Low	Low
T2	High	High

Table-2 The chiller and cooling tower locations

Component	Location
Chiller	Basement
Cooling tower	Rooftop

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

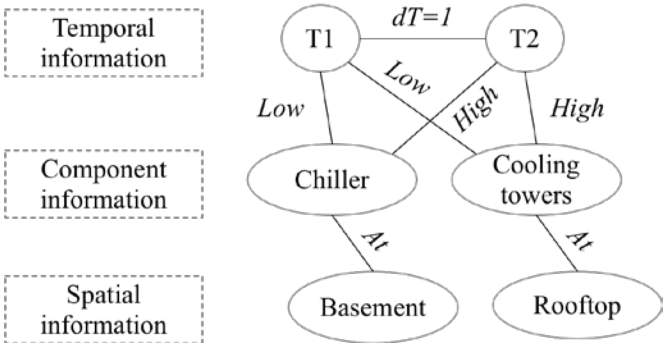

Fig. 1 The graph for information integration and representation

2.2 A brief overview of graph mining techniques

Graph mining aims to extract insightful and actionable knowledge from graphs. It can be regarded as a specialization of data mining, where specific techniques are needed to handle graph data other than tabular data [34]. Various graph mining techniques have been proposed to extract predictive and descriptive knowledge from graphs. In terms of graph-based predictive modelling, there are two main tasks, i.e., link prediction and graph classification. Link prediction aims to predict whether there will be a link between two vertices or not [35]. It has been widely used in the field of social network analyses to infer the relationships among people. The main idea is to develop a logistic model based on the proximities of different vertices, which are calculated based on the graph topology [35]. The second is graph classification, which aims to predict a label to a graph based on its structures. For instance, the molecular structure of various chemical compounds can be naturally represented as graphs and a graph-based classification model can be developed to predict if the chemical compound is toxic or not [36]. The main technical challenge is to accurately evaluate proximities among graphs, which also serves as the basis for graph-based descriptive knowledge discovery, e.g., graph clustering and graph-based anomaly detection [36].

Conventional data representations usually use a feature vector to describe an observation and therefore, the proximities between two observations can be calculated using distance measures, e.g., the Euclidean distance. Due to the intrinsic structures of graph data, specific measures are needed for proximity evaluation. There are two

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

general approaches for graph-based proximity evaluation. The first is to calculate graph-level indices as features for each graph and conventional distance measures can then be applied for proximity evaluation. The most commonly used graph-level indices include the total number of vertices and edges, the mean degree of graph vertices, the graph density and graph diameter [37]. Graph-level indices are easy for implementation, yet they cannot accurately describe the graph structures and are not applicable for labelled graphs. The other approach is based on the concept of subgraph mining [38]. In such a case, the proximity between two graphs is evaluated based on their common subgraphs or substructures. Such approach is compatible with labeled graphs and can better preserve the topological characteristics of graphs. Similar to frequent itemset mining in analyzing tabular data, frequent subgraph mining is the essence for most graph data analytics [39]. The concept of frequent subgraph mining is introduced in the following section.

2.3 Frequent subgraph mining

Frequent subgraph mining (FSM) aims to discover frequent subgraphs whose occurrences or frequencies exceed a minimum threshold (i.e., minimum support threshold). FSM typically works with undirected and labelled graphs. The frequent subgraphs discovered can be directly used to represent the frequent patterns, e.g., finding the common substructures of chemical compounds and identifying frequent behavioral patterns of terrorist attacks [39]. They can also be used as the basis for other graph mining tasks, such as graph classification and graph clustering [40].

In general, FSM algorithms, which take a set of graphs as input, can be classified using two criteria, i.e., (1) whether it adopts an exact or inexact search strategy; (2) whether the search strategy is breadth-first or depth-first [39]. Inexact search FSM algorithms, such as SUBDUE [41] and CREW [42], adopt approximated measures to compare two graphs. The mining efficiency is typically higher at the expense of not finding all frequent subgraphs. Exact FSM algorithms are more commonly used as they can discover all frequent subgraphs. Such algorithms can be further classified based on

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

whether the search strategies is breadth-first or depth-first. The depth-first search strategy is typically more computationally efficient, such as *MoFa* [43], *gSpan* [44], *FFSM* [45] and *GASTON* [46]. A recent study compared the performance of these four algorithms, indicating that *gSpan* is better in terms of computation time and memory usage [47].

It should be noted that the number of frequent subgraphs discovered can be very large while the majority of them are redundant. A subgraph is redundant if there is a supergraph with the same or larger number of appearances (i.e., the support). To overcome this challenge, Yan and Han proposed an algorithm called *CloseGraph* to mine closed frequent graphs based on the *gSpan* algorithm [48]. A subgraph is closed if there exists no supergraph with the same or larger supports. It was shown that *CloseGraph* could dramatically reduce the number of redundant subgraphs and enhance the mining efficiency. In this study, the *CloseGraph* is adopted for frequent subgraph mining. The algorithm takes a set of graphs as input. A minimum support threshold, which is used to evaluate whether a subgraph is frequent or not, should be pre-defined for implementations.

3. Research Methodology

3.1 Research outline

The research methodology is developed to tackle challenges associated in analyzing massive building operational data. As shown in Fig. 2, it consists of four phases, i.e., data exploration, graph generation, knowledge discovery and knowledge post-mining. The first phase, i.e., data exploration, aims to reveal the general building operational patterns and data structures through data-driven approaches. The insights obtained are used as the basis for further analyses. At the second phase, graph generation methods are developed to transform relational building operational data into graphs. Compared with conventional data formats, graphs are suitable for integrating and representing different types of information and therefore, are more promising for high-level knowledge discovery and visualization. At the third phase, frequent subgraph mining is

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

adopted to extract statistically significant subgraphs from the graph data set. Post-mining methods are then developed to transform raw data analysis results into applicable insights for building energy management.

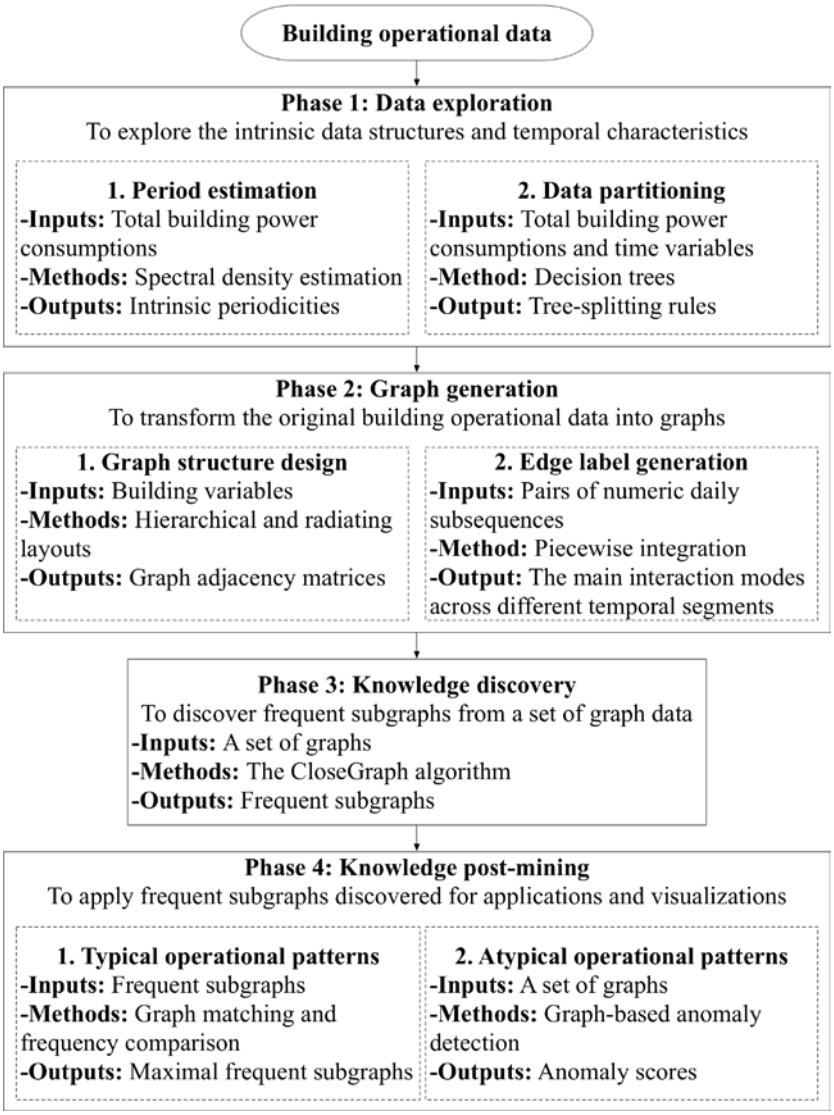

Fig. 2 Research outline

3.2 Data exploration

Building operational data have two prominent features. Firstly, building operational data are in essence time series data. The periodicity is one of the most important characteristics of time series and can be used to determine the data formats for further analyses. Hence, the first task is to identify the intrinsic periodicities in building operational data. Secondly, building operational data are highly complex due to the existence of multiple operating modes and the dynamic interactions among various

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

building variables. To enhance the reliability and sensitivity of data analyses, it is necessarily to divide the whole data according to different operating modes for separate knowledge discovery. Hence, the second task in data exploration is to identify typical operating modes for data partitioning.

More specifically, the spectral density estimation method is adopted to identify the intrinsic periodicities in building operational data. The building power consumption is selected for analysis as it is one of the major concerns of buildings and can reflect typical building operating patterns. The fast Fourier transformation is used to calculate the periodogram of the total building power consumption, based on which data smoothing is applied using a series of moving average smoothers. The intrinsic periodicities can then be discovered by identifying peaks in spectral densities. Once the most dominant periodicities in building operations are identified, the time series data can be divided into shorter subsequences for detailed analyses.

The decision tree method is used to identify general building operating modes for data partitioning. It is selected due to the high interpretability of the resulting model. Time variables (e.g., *Year*, *Month*, *Day Type* and *Hour*) are selected as model inputs, while the total building power consumption is selected as the model output. The resulting model is able to describe the general operating modes in terms of time variables and the tree-splitting criteria are utilized for data partitioning. It is worth mentioning that the time variables are categorical variables with different numbers of possible values, e.g., *Month* has 12 unique values while *Hour* has 24 unique values. To avoid the selection bias during tree model development, the unconditional inference tree algorithm [49] is adopted in this study.

3.3 Graph generation

In this study, a variable-based method is proposed to transform relational building operational data into graphs. The method is developed with two considerations, i.e., to minimize the computational burdens of data analysis and ensure the compatibility with frequent subgraph mining algorithms. The former requires the number of vertices and

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

edges used for information representation is minimized, while the latter requires the graph to be labelled and interconnected (i.e., there is always a path from one vertex to another). To meet these requirements, a radiating graph layout and a unique edge labelling scheme is developed for graph generation.

More specifically, a radiating graph layout is designed to preserve the hierarchical information among building variables. In this study, each building variable is denoted as a vertex. The radiating layout can be used to represent the hierarchical information in building services systems. The center vertex represents building-level variables, such as the total building cooling load or power consumption. The first outer layer of vertices denotes system-level or subsystem-level variables, such as the chiller plant or a space location. The second outer layer of vertices denotes the component-level variables, such as individual chillers and water pumps. A third outer layer can be developed to represent the physical operating parameters of each component, e.g., the supplied and returned chilled water temperatures of each individual chiller.

To describe the interactions among building variables, a unique edge labelling scheme is proposed. An intuitive metric to describe the interaction between two variables is the pairwise correlation. However, it cannot reflect the absolute values of each variable. For instance, a high correlation between two components' power consumptions cannot reflect their actual values, e.g., either *Low* or *High*. To tackle these limitations, a three-step piecewise aggregation method is developed for edge labelling: (1) Transform numeric time series subsequences into categorical values using data discretization techniques, e.g., equal-width or equal-frequency binning; (2) Divide each subsequence into k non-overlapping temporal segments based on the periodicities discovered at the data exploration stage; (3) Extract the most frequent interaction modes between two variables in each temporal segment. An interaction mode is defined as a character string containing the categorical values of both variables. For instance, if variable A is *Low* and variable B is *High*, the interaction mode is defined as {*Low*, *High*}. The interaction mode can be further transformed into integers for the ease of notation. Fig. 3 and Table-

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

3 present an illustration of edge labelling, assuming that each numeric subsequence is divided into three temporal segments and categorized into two levels, i.e., *Low* and *High*. The edge label is "1-2-1", indicating that the most frequent interactions are {*Low*, *Low*}, {*Low*, *High*} and {*Low*, *Low*} in three temporal segments respectively.

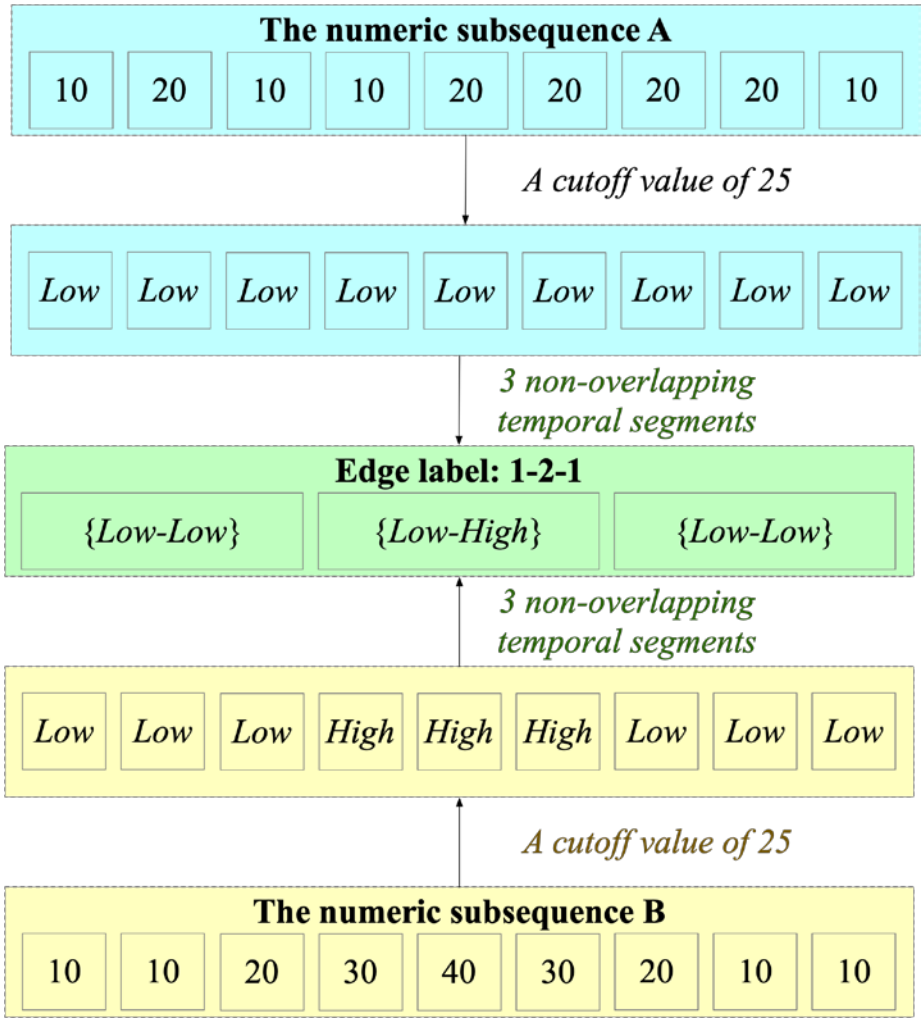

Fig. 3 An illustration of edge labelling

Table-3 An example notation scheme for interaction mode representation

Inner variable A	Outer variable B	Interaction mode	Notation
Low	Low	{Low, Low}	1
Low	High	{Low, High}	2
High	Low	{High, Low}	3
High	High	{High, High}	4

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

3.4 The knowledge post-mining methods

Once the graph data set is constructed, the frequent subgraph mining is adopted to discover frequent subgraphs. The *CloseGraph* algorithm is selected as it can greatly reduce the number of redundant subgraphs. Two post-mining methods are developed to ensure the efficiency in knowledge summarization and application, i.e., maximal frequent subgraph identification and graph-based anomaly detection.

Firstly, given a set of frequent subgraph F, a traversal operation is performed to find the set of maximal frequent subgraphs M. Each graph in M is then a maximal frequent subgraph with no supergraph identified as frequent. The maximal frequent subgraphs are used to represent typical operation patterns.

Secondly, anomaly detection is performed based on the set of maximal frequent subgraphs. The general idea is that a graph is more likely to be an anomaly if it differs from the frequent subgraphs discovered. Assuming that Y frequent subgraphs are discovered based on X graphs, the anomaly detection method will output an anomaly score for each of the X graphs. For a given graph G_i, the anomaly score is defined as $A_i = \frac{1}{Y}\sum_{j=1}^{Y}\frac{D_{i,j}}{N_{s,j}}$, where $D_{i,j}$ is the minimal number of differences in vertices and edges between G_i and the j^{th} frequent subgraph, $N_{s,j}$ is the total number of vertices and edges in the j^{th} frequent subgraph. Fig. 4 presents an illustration of graph-based anomaly detection. The minimal number of differences is two, i.e., one in vertices and one in edge labels. Hence, the anomaly score is $\frac{2}{7}$, where 2 is the total number of differences and 7 is the total number of vertices and edges in the frequent subgraph.

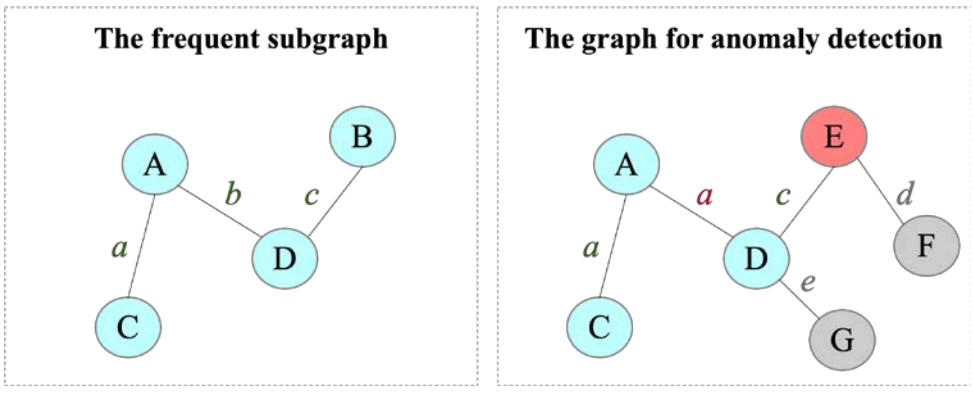

Fig. 4 An illustration of graph-based anomaly detection

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

4. Mining actual building operational data

4.1 Building and data descriptions

The building operational data used in this study were retrieved from a public exhibition building in Hong Kong. The building has a total site area of 14,700m^2. The majority of the site is a landscaped area for public use. An eco-café and a small shop are located in the landscaped area. The main component in the site is a 3-storey building with a footprint of 1,400m^2. It consists of an exhibition area, an eco-home, an eco-office and a multi-purpose hall. Several passive design features have been integrated for energy saving, such as cross-ventilated layout, high performance glazing, light pipes and earth cooling tube. The active systems integrated include high-volume-low-speed fans, high temperature cooling system, intelligent lighting management, absorption chiller, photovoltaic panels and bio-diesel tri-generation systems. Pre-defined control logics have been developed for managing renewable energy, e.g., utilizing daylighting during daily operations and controlling photovoltaic panels for electricity generation. The estimated energy use for the building and the landscape area is around 116MWh and 15MWh per year respectively. The major energy generation components are the biodiesel tri-generator and PV panels and their estimated energy outputs are 143MWh and 87MWh per year respectively.

A building automation system has been installed to monitor and control the building operational performance over various subsystems. A data set with one-year operational data was adopted for analysis. The data contain hourly cooling load demands and all the power measurements of major space areas or services components. The data set has 8,304 observations and 38 variables, such as the *Year*, *Month*, *Day*, *Hour*, *Day Type*, the power consumptions of three water-cooled chillers (*WCC-1 to 3*), four chilled water pumps (*CHWP-1 to 4*), three condenser water pumps (*CDWP-1 to 3*), three cooling towers (*CT-1 to 3*), five air-handling units (*AHU-1 to 5*), one primary air-handling unit (*PAU*), the power consumptions of outdoor landscape lighting (*LandLight*), the normal power and lighting consumptions of the eco-areas (*Eco-office* and *Eco-café*), basement

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

area (*Base*), G/F common area (*GF*), multi-purpose room (*MPR*) and mezzanine area (*Mezz*). The open-source software *R* [50] and its associated packages *igraph* [51] and *ggraph* [52] were used for graph generations and data visualizations. The software *ParSeMis* was used for frequent subgraph mining [53].

4.2 Insights obtained by data exploration

The non-parametric spectral density estimation method was used to identify intrinsic periodicities in total building power consumptions. As shown in Fig. 5, the frequency corresponding to the highest spectral density peak is 0.042, indicating a period of 23.8 (i.e., $\frac{1}{0.042}$). Considering that the data were collected at hourly basis, it indicates a significant daily periodicity.

The other task in data exploration is to extract insights for data partitioning, which ensures the sensitivity and reliability of in-depth analyses. As described in Section 3.2, the decision tree method was used to capture the relationship between the total building power consumption and time variables (i.e., *Year*, *Month*, *Day*, *Hour*, and *Day Type*). The resulting model is shown in Fig. 6. Three variables, i.e., *Day Type*, *Hour* and *Month* were automatically selected as the splitting variables. The root node selects the *Hour* as the splitting variable. The splitting criteria are {9 to 19} and the others. Such criteria are in accordance with domain expertise, as they correspond to office and non-office hours of this building. Node 3 selects the *Day Type* as the splitting variable. The data are partitioned based on whether they were collected on Wednesday and Sundays or not. It turns out that the building is closed for exhibition on Wednesday and Sundays and therefore, the splitting criteria are in accordance with working and non-working days. Nodes 4 and 7 both select the *Month* as the splitting variable and the splitting criteria are in accordance with the hot and cold seasons in Hong Kong.

The insights obtained were used in three ways. Firstly, the spectral density analysis suggests a significant daily periodicity and hence, the building operational data were divided into daily subsequences. Secondly, as summarized in Table-4, daily subsequences were partitioned into four groups for separate knowledge discovery

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

according to the *Month* and *Day Type*. Thirdly, each daily subsequence was divided into three temporal segments according to *Hour* when generating edge labels for graphs, i.e., {0 to 8}, {9 to 19} and {20 to 23}.

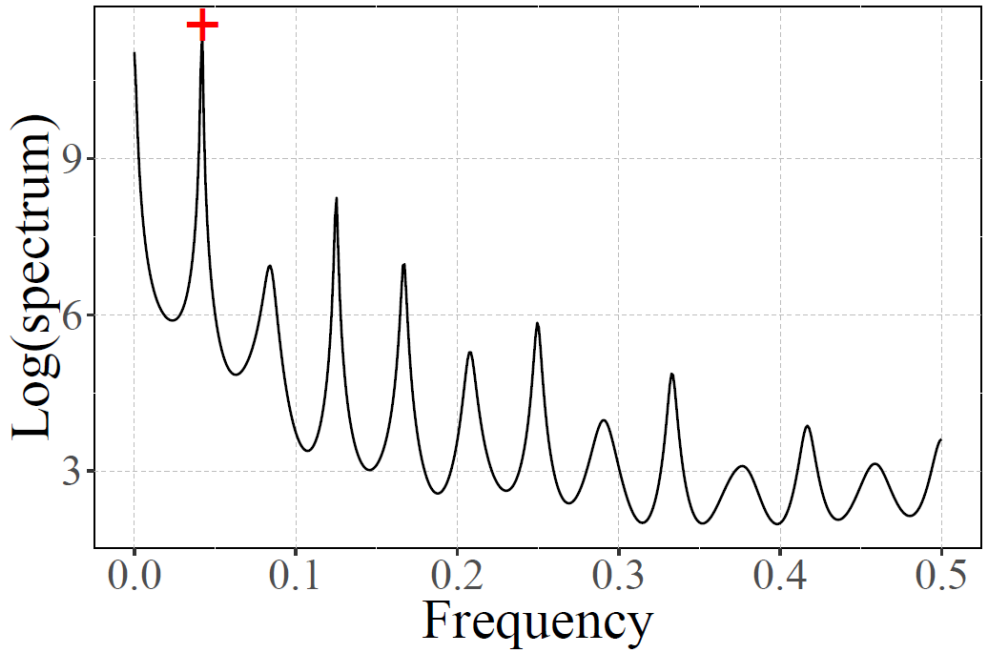

Fig. 5 Spectrum density estimation for the time series of building cooling load

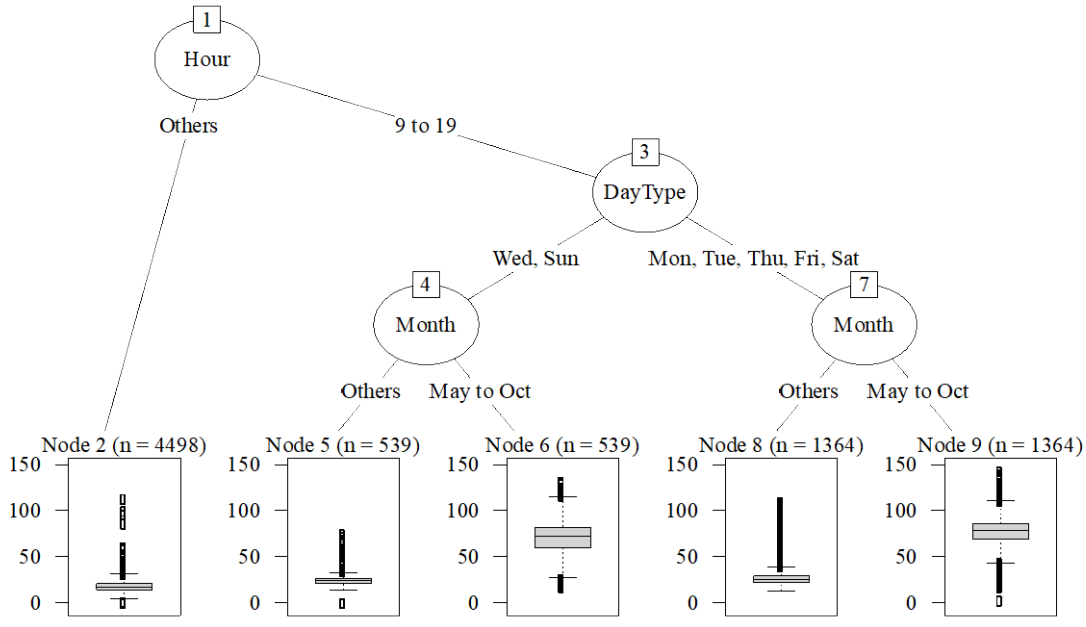

Fig. 6 The decision tree model developed for data exploration

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

Table-4 A summary on data partitioning

Groups	Month	Day Type	No. of subsequences
1	{1,2,3,4,11,12}	{Mon, Tue, Thu, Fri, Sat}	124
2	{1,2,3,4,11,12}	{Wed, Sun}	49
3	{5,6,7,8,9,10}	{Mon, Tue, Thu, Fri, Sat}	124
4	{5,6,7,8,9,10}	{Wed, Sun}	49

4.3 Transforming building operational data into daily graphs

The graph generation method proposed in Section 3.3 was used to transform building operational data into daily graphs. The power consumption data for different space areas and components were divided into daily subsequences, based on which the pairwise high-level temporal interaction modes were extracted for graph generation. The graph to be generated has a radiating layout to preserve the hierarchical information among building variables. In this study, the center vertex denotes the total building cooling load demand. It is connected with system-level vertices, which represent the power consumptions of different spaces and HVAC subsystems. The system-level vertices are connected with component-level vertices, which represent the power consumption of individual HVAC components.

To create edge labels, the power consumption data were firstly preprocessed using the max-min normalization ($x' = \frac{x-\min(x)}{\max(x)-\min(x)}$) and then discretized into three levels, i.e., *Idle*, *Low* and *High*. In this study, the power consumption data were categorized as *Idle* if their normalized values are smaller than a threshold, i.e., 0.05. The equal width binning method was then applied to generate *Lows* and *Highs* based on a cutoff value of 0.5. As indicated in Section 4.2, three dominant operation modes can be extracted for each daily subsequence according to three temporal segments, i.e., {0 to 8}, {9 to 19} and {20 to 23}. The pairwise interaction modes between any pair of daily subsequences can be created according to the notation scheme shown in Table-5.

Fig. 7 illustrates an example daily graph generated on July 15, 2013 (Tuesday). The center vertex, the system-level vertices and the component-level vertices are shown in

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

red, blue and green respectively. Each vertex represents a building variable and the edges are labelled accordingly. For instance, the edge label between the *Load* and *Pwr_WCC* is 1-5-1, which means that the dominant interaction modes across three temporal segments are {*Idle, Idle*}, {*Low, Low*} and {*Idle, Idle*} respectively. Such graph has the ability to integrate and represent hierarchical information and temporal interactions among building variables. It serves as a new type of information carrier, based on which useful insights can be discovered for building energy management.

Table-5 The notation scheme used for interaction mode representation

Variable A	Variable B	Interaction mode	Notation
Idle	Idle	{Idle, Idle}	1
Idle	Low	{Idle, Low}	2
Idle	High	{Idle, High}	3
Low	Idle	{Low, Idle}	4
Low	Low	{Low, Low}	5
Low	High	{Low, High}	6
High	Idle	{High, Idle}	7
High	Low	{High, Low}	8
High	High	{High, High}	9

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

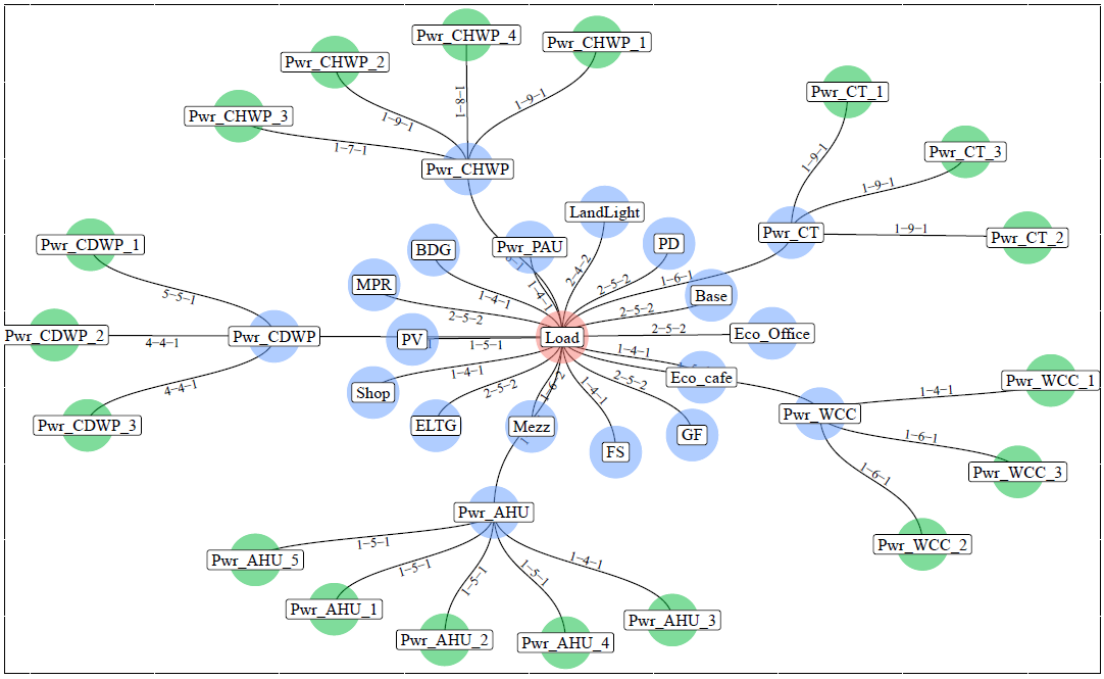

Fig. 7 An example daily graph generated on July 16, 2013 (Tuesday)

4.4 Discovering typical building operational patterns

As described in Section 4.2, daily graphs were divided into four groups based on *Month* and *Day Type* for separate analysis. The frequent subgraph mining was applied with a minimal support threshold of 40%. As introduced in Section 3.4, a post-mining method has been developed to identify maximal frequent subgraphs with the aim of alleviating the burden of manual inspection. The numbers of maximal frequent subgraphs discovered for each data group are summarized in Table-6. It is observed that the numbers of maximal frequent subgraphs discovered are larger during hot seasons and working days, while smaller during cold seasons and non-working days. In addition, the relative sizes of maximal frequent subgraphs discovered in each data group are calculated and visualized in Fig. 8. The relative size is defined as the ratio between the size of frequent subgraphs and the complete daily graphs. It is shown that the relative sizes of frequent subgraphs in the cold seasons (i.e., Groups 1 and 2) are much larger than those in the hot seasons. This is expected as the smaller the cooling loads, the fewer variations in the operating conditions of the HVAC system and hence, the sizes of frequent subgraphs becomes larger.

The maximal frequent subgraphs discovered can be used to describe the typical building

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

operational patterns and high-level interactions among building variables. To illustrate, Figs. 9 and 10 present examples of frequent subgraphs discovered in Groups 1 and 2. It is shown that the total building cooling load is *Idle*, and the HVAC system is completely switched off. The other building variables generally fall into the categories of *Idle* or *Low*. One noticeable difference is the edge label between *Load* and *Eco-office*, which is "2-2-2" on working days and "1-1-1" on non-working days in cold seasons. It indicates that the *Eco-office* is vacant during non-working days and hence, its normal power and lighting consumptions are categorized as *Idle*. Figs. 11 and 12 present examples of frequent subgraphs discovered in Groups 3 and 4. It is shown that the HVAC system is switch-on during office hours. Similarly, the interaction modes between *Load* and *Eco-office* are different during working and non-working days.

Table-6 The number of maximal frequent subgraphs discovered

Groups	Month	Day Type	No. of maximal frequent subgraphs
1	{1,2,3,4,11,12}	{Mon, Tue, Thu, Fri, Sat}	11
2	{1,2,3,4,11,12}	{Wed, Sun}	10
3	{5,6,7,8,9,10}	{Mon, Tue, Thu, Fri, Sat}	23
4	{5,6,7,8,9,10}	{Wed, Sun}	16

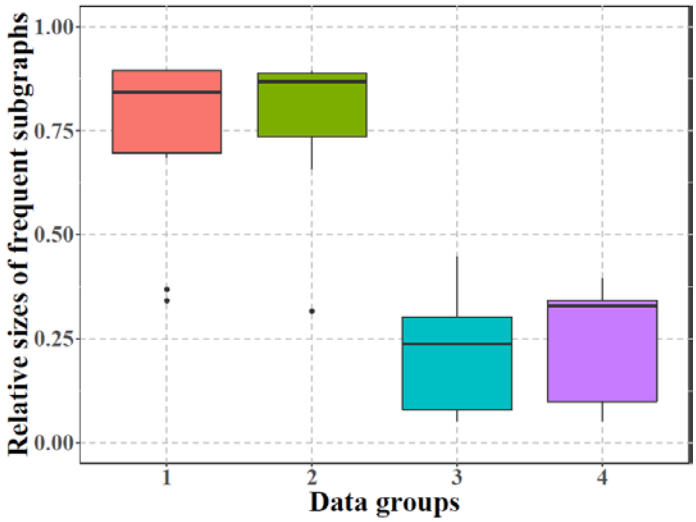

Fig. 8 The relative sizes of maximal frequent subgraphs in different data groups

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

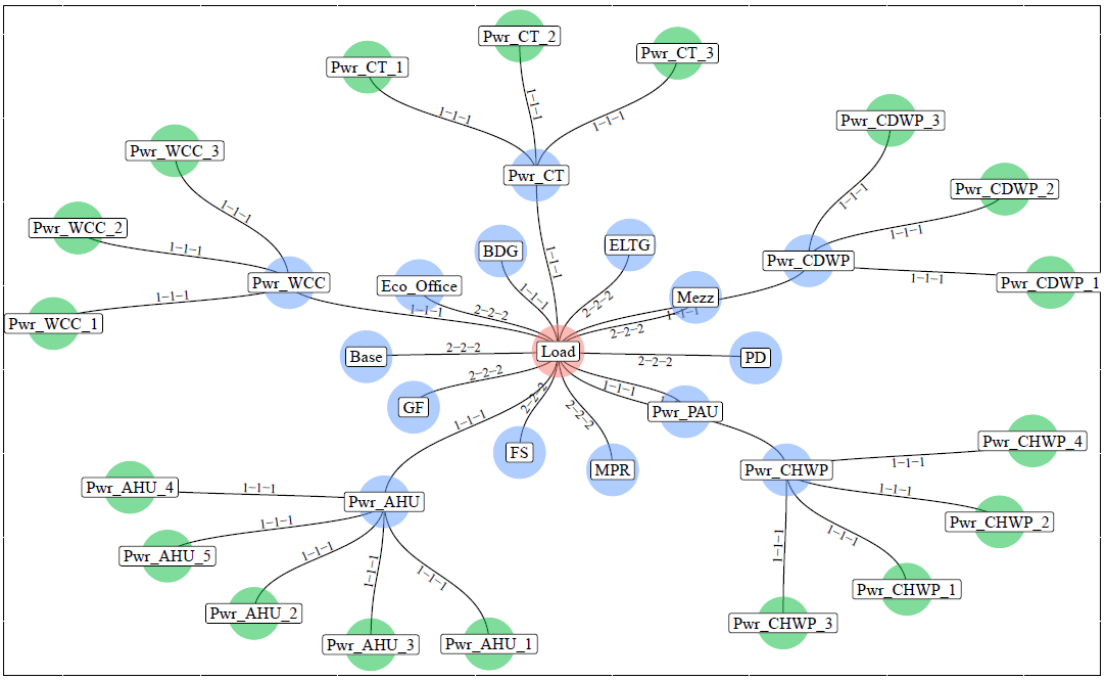

Fig. 9 An example of maximal frequent subgraph discovered in Group 1

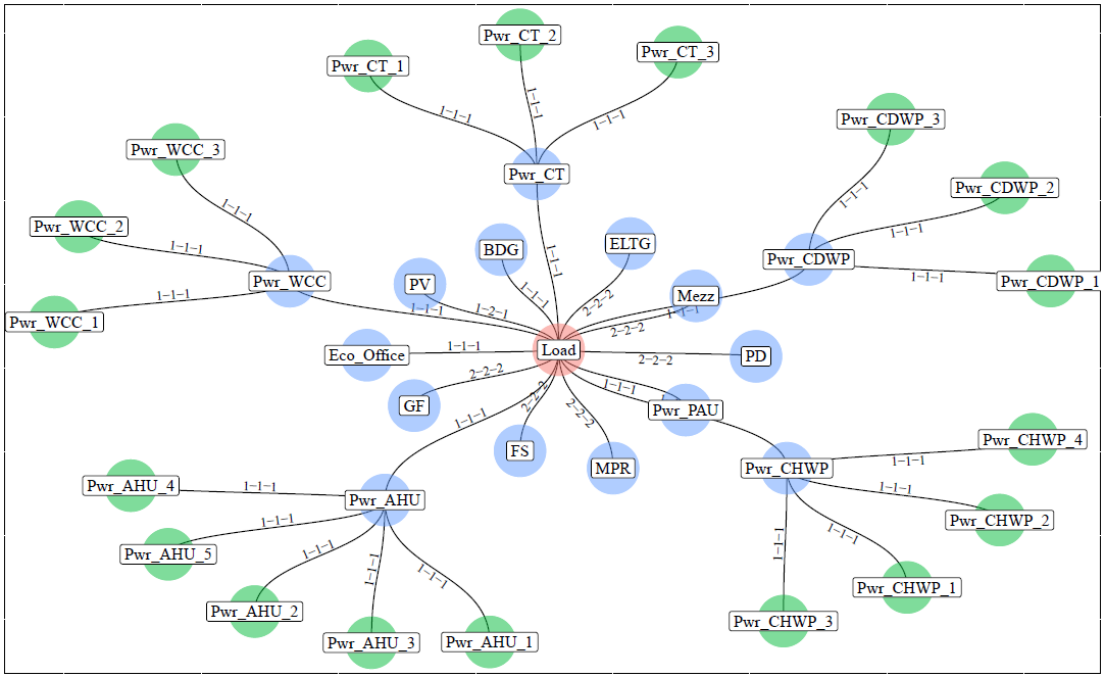

Fig. 10 An example of maximal frequent subgraph discovered in Group 2

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

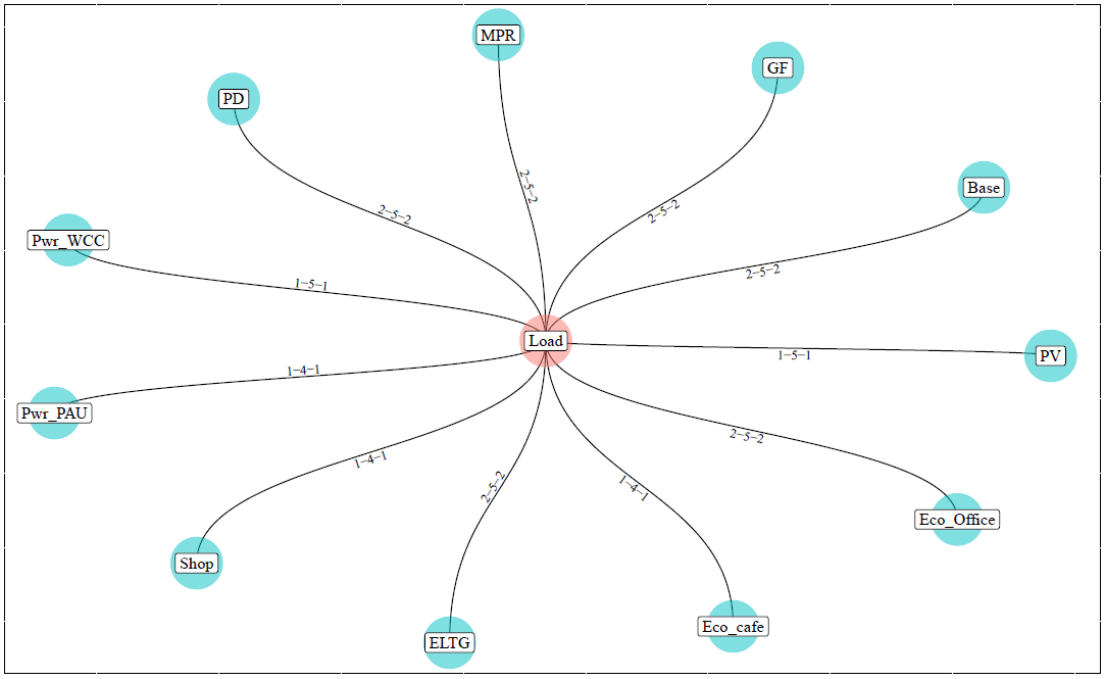

Fig. 11 An example of maximal frequent subgraph discovered in Group 3

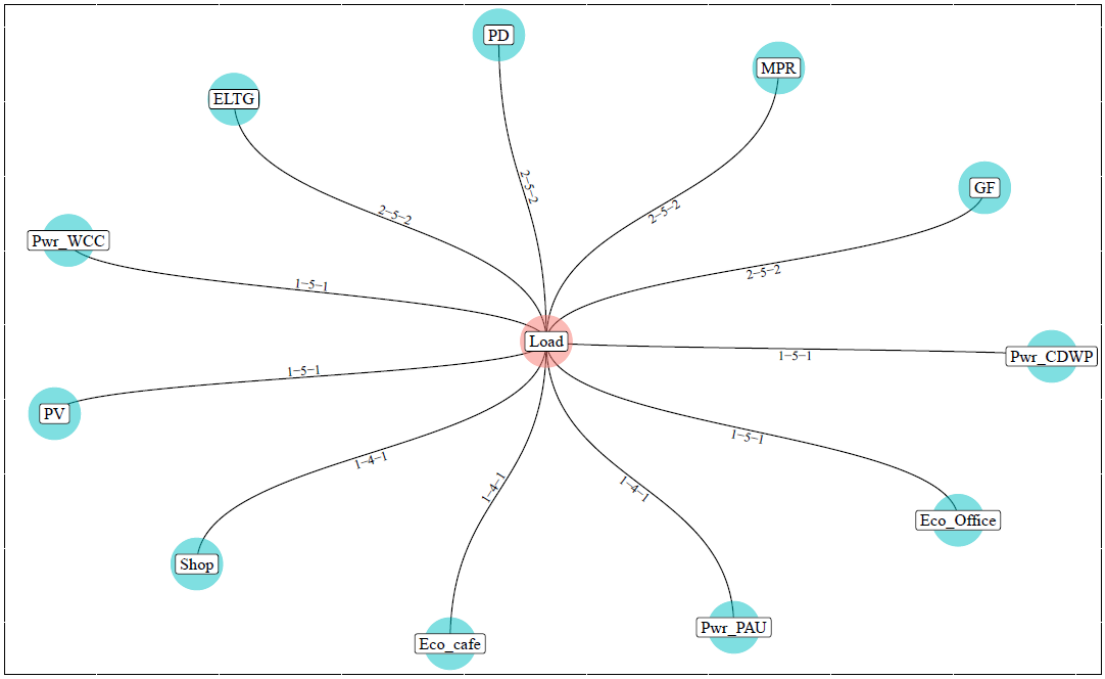

Fig. 12 An example of maximal frequent subgraph discovered in Group 4

4.5 Discovering atypical building operational patterns

As introduced in Section 3.4, a graph-based anomaly detection method has been developed for knowledge post-mining. For each daily graph, an anomaly score is calculated based on the maximal frequent subgraphs discovered. Such scores can be

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

used to facilitate building operation staffs to quickly identify potential anomalies, while atypical operational patterns can be readily presented as graphs for the ease of interpretations.

An example of the anomaly detection results is shown in Fig. 13 and the maximal frequent subgraph used for comparison is shown in Fig. 14. The graph vertices and edges are colored in green and red, corresponding to matched and non-matched parts. The other vertices and edges are colored in grey, indicating that they are not used for comparison with regard to the frequent subgraph considered. It is shown that the mismatches take place between *Pwr_AHU*, *Pwr_AHU_1*, *4* and *5*. As shown in Fig. 14, the edge labels between the total power consumptions of AHU and power consumptions of *AHU-1*, *4* and *5* should be "1-5-1", indicating that the dominant interaction modes should be {*Idle*, *Idle*} during 0 a.m. to 8 a.m., {*Low*, *Low*} during 9 a.m. to 7 p.m., and {*Idle*, *Idle*} during 8 p.m. to 11 p.m. However, as shown in Fig. 13, the interaction modes change to {*Idle*, *Idle*} during the office hours for *AHU-4* and *5*, while {*Low*, *High*} for *AHU-1*. In such a case, the atypical operation identified is an infrequent but normal operation, as it may due to the changes of equipment working schedules.

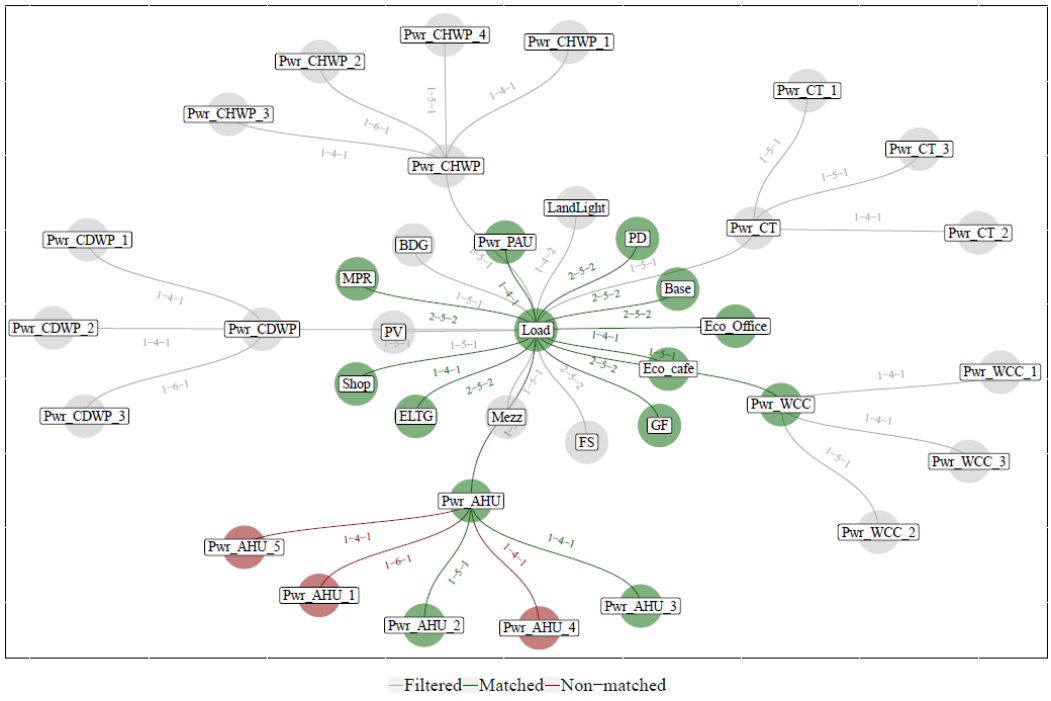

Fig. 13 Case 1: The daily graph on October 14, 2013 (Monday)

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

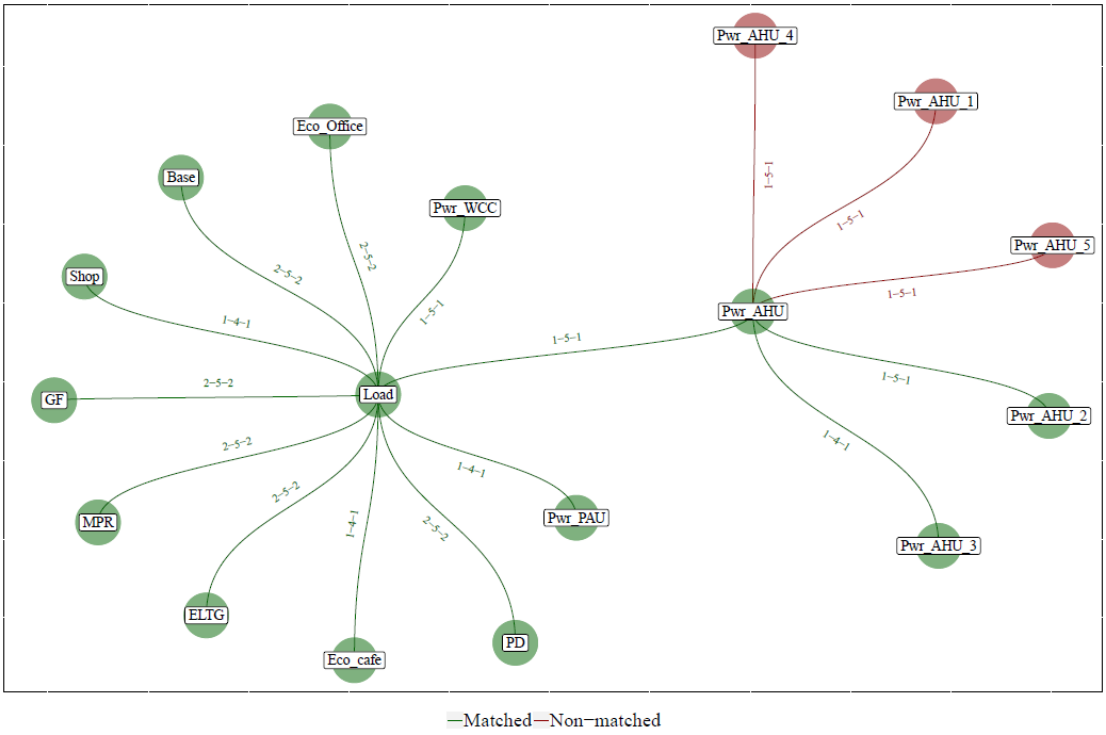

Fig. 14 The reference frequent subgraph used for the anomaly detection in Case 1

Figs. 15 and 16 present another anomaly detection example. It is observed that the anomaly comes from the interactions between total building cooling load and the power consumption of the basement. Conventionally, the interaction modes across three temporal segments should be {*Idle*, *Low*}. However, the interaction modes changed to {*Idle*, *Idle*} in the last temporal segments in Fig. 15. Further investigation reveals that the atypical graph represents the building operations on December 26, 2013, which is a public holiday in Hong Kong. After consulting with the operation staff, it is found out that normally, the lighting in basement should be switched on at non-office hours. It was manually switched off on that day as the next day is also a public holiday and there was no working plan in that area.

As shown in Figs. 17 and 18, the daily graph on October 30, 2013 was identified as an anomaly due to interactions between *Load* and *LandLight*, *Pwr_AHU* and *Pwr_AHU_2*. More specifically, the normal interaction modes between the total building cooling load and landscape lighting should be “2-4-2”, representing {*Idle*, *Low*}, {*Low*, *Low*} and {*Idle*, *Low*} respectively. Nevertheless, the interaction mode between 0 a.m. to 8 a.m. changed to {*Idle*, *Idle*} on October 30, 2013, indicating that there is no landscape

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

lighting at all. After consulting with the building operation staff, it is found that the landscape lighting should be automatically switched on from 7 p.m. to 7 a.m. Therefore, the anomaly identified might be caused by manual faults or maintenance.

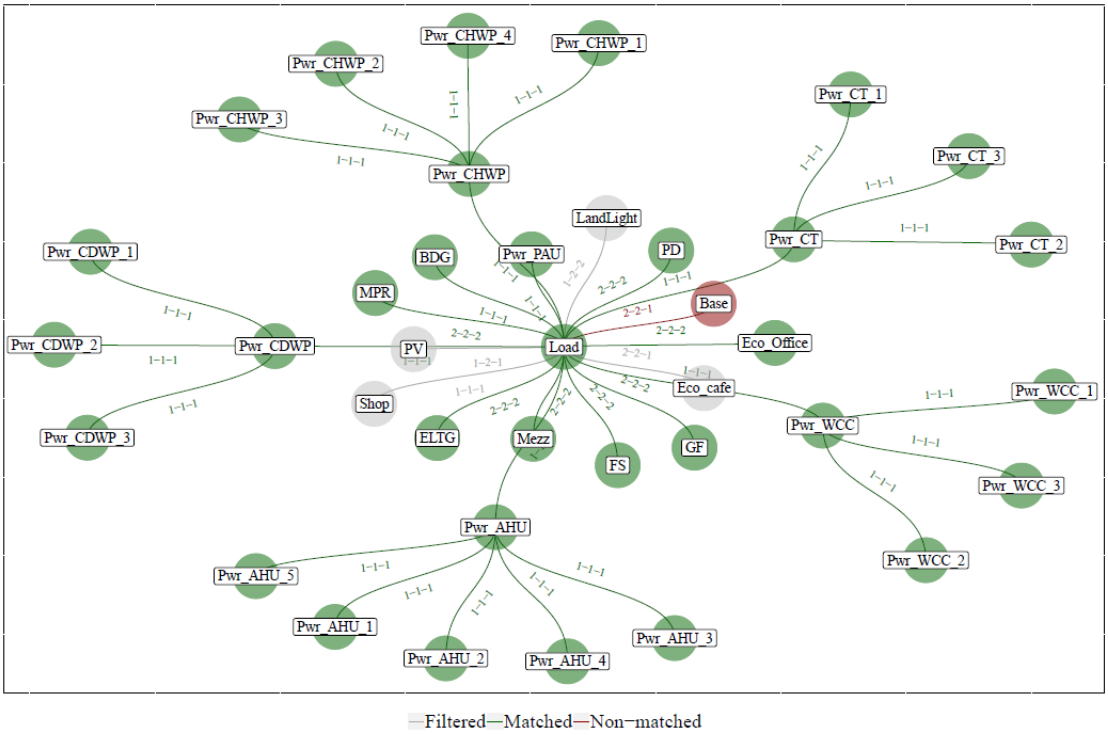

Fig. 15 Case 2: The daily graph on December 26, 2013 (Thursday)

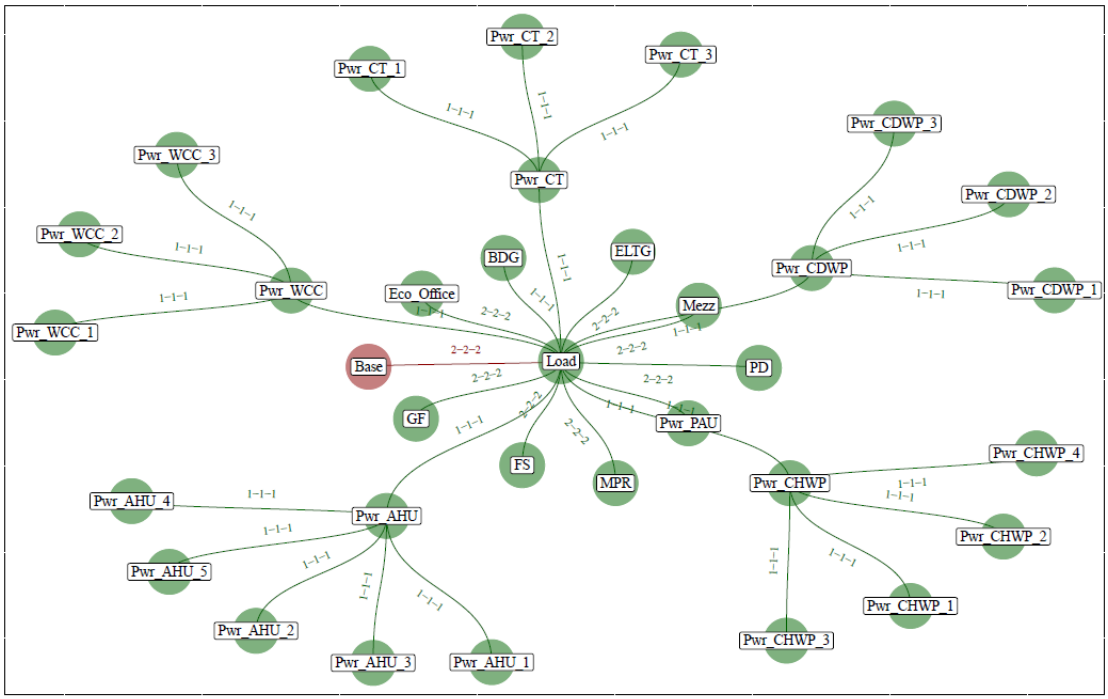

Fig. 16 The reference frequent subgraph used for the anomaly detection in Case 2

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

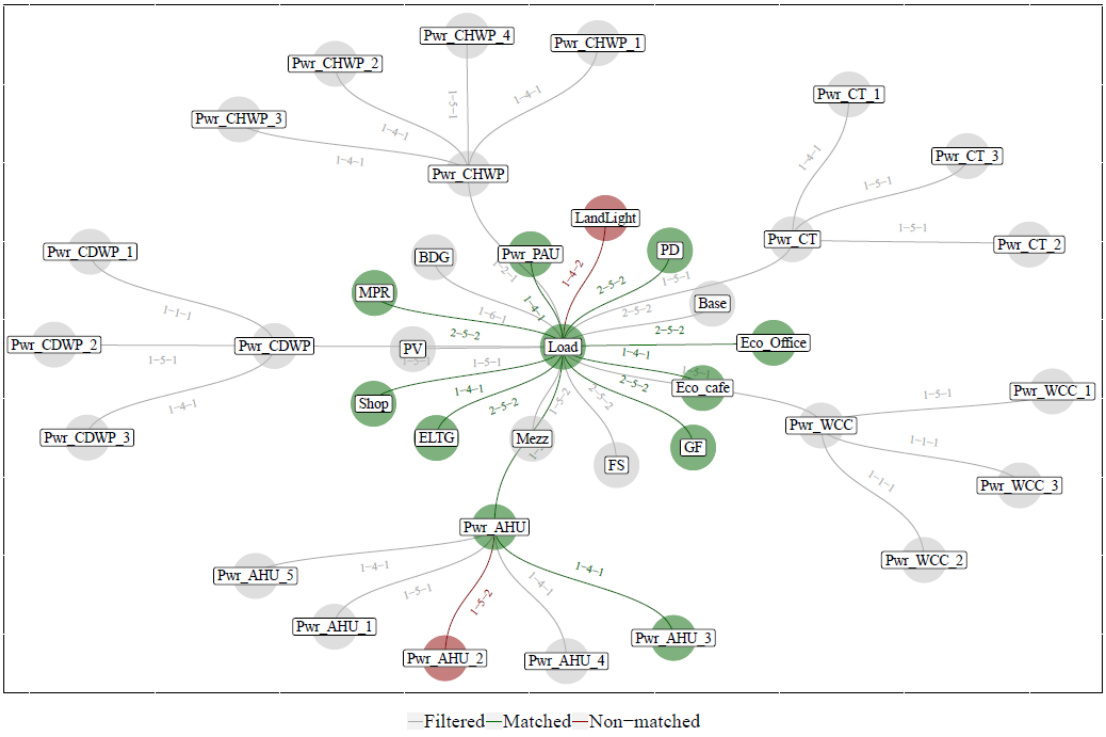

Fig. 17 Case 3: The daily graph on October 30, 2013 (Wednesday)

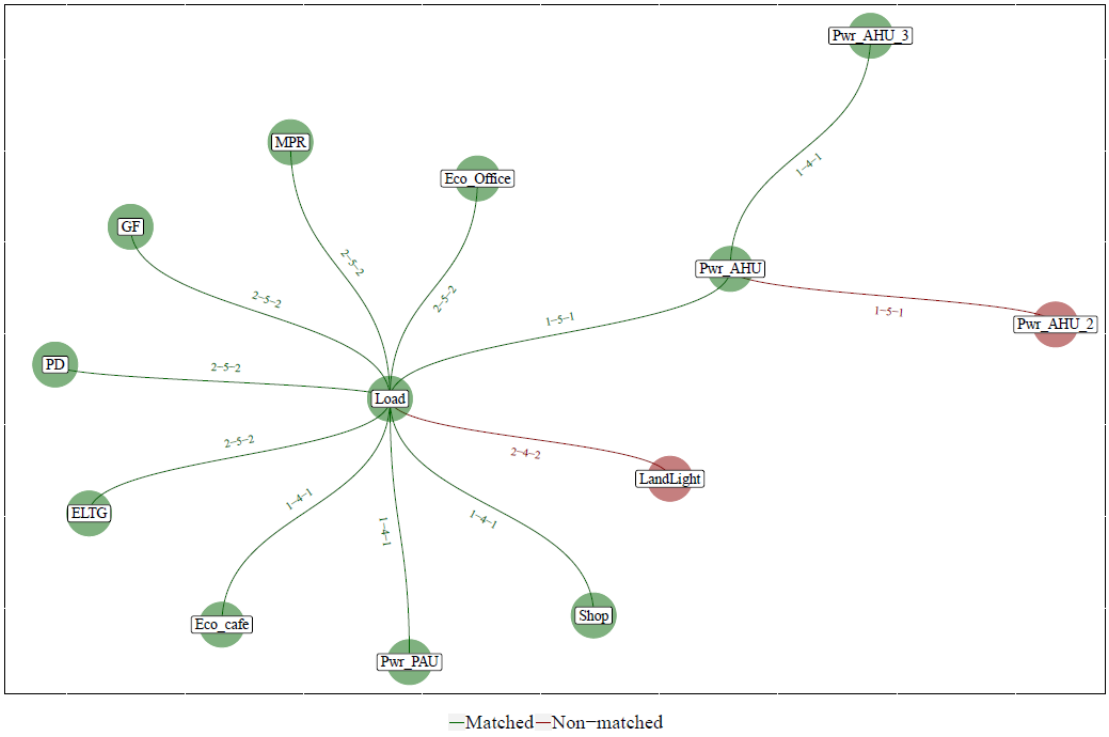

Fig. 18 The reference frequent subgraph used for the anomaly detection in Case 3

5. Conclusions

Building are becoming increasingly information-intensive. It can be foreseen that more types of information will be collected and available for data analysis. Therefore, advanced data analytics, which are capable of integrating and representing complicated

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

information, should be developed to fully embrace the era of big data. This study proposes a graph-based methodology to integrate, represent and discover high-level knowledge from original building operational data. The methodology is developed to maximize the data analysis efficiency while minimizing the computational burdens. To summarize, a variable-based method is proposed to transform relational building operational data into graphs. A radiating layout is adopted to preserve the hierarchical information among building variables. The temporal interactions among building variables are represented as edge labels. The frequent subgraph mining is adopted to discover statistically significant building operation patterns. Specific post-mining methods have been proposed for knowledge interpretation, summarization and applications. The methodology has been applied to analyze actual building operational data retrieved from a public building in Hong Kong. The research results validate the potential of graph-based methods in characterizing building operational patterns and identifying atypical operations.

It should be noted that the building operational data analyzed were collected at hourly basis. As shown by the spectral density estimation results, such temporal resolution is capable of describing interactions with a daily periodicity. In-depth analyses can be performed if the data were collected at higher temporal resolutions, e.g., discovering the high-level temporal interactions among building variables during the chiller stage-on or stage-off periods. In such a case, more detailed or accurate interactions among building variables can be captured, while at the cost of increasing computational burdens. Future studies will be carried out to develop advanced edge labelling schemes to balance the trade-off between information loss and computational burdens in analyzing building operational data with higher temporal resolutions.

Acknowledgement

The authors gratefully acknowledge the support of this research by the Natural Science Foundation of Guangdong Province, China (No. 2018A030310543), the Philosophical

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

and Social Science Program of Guangdong Province, China (No. GD18YGL07), and the National Taipei University of Technology-Shenzhen University Joint Research Program (No. 2019003).

References

[1] Dalene F. Technology and information management for low-carbon building. J Renew Sust Energ 2012; 4-041402. doi: 10.1063/1.3694120.

[2] Waide P, Ure J, Karagianni N, Smith G, Bordass B. The scope for energy and CO2 savings in the EU through the use of building automation technology. Final Report for the European Copper Institute. August 10, 2013.

[3] Amasyali K, El-Gohary NM. A review of data-driven building energy consumption prediction studies. Renew Sustain Energy Rev 2018; 81: 1192-205.

[4] Fan C, Wang JY, Gang WJ, Li SH. Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. Appl Energy 2019; 236: 700-10.

[5] Fan C, Xiao F, Zhao Y. A short-term building cooling load prediction method using deep learning algorithms. Appl Energy 2017; 195: 222-33.

[6] Fan C, Xiao F, Wang SW. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl Energy 2014; 127: 1–10.

[7] Chou JS, Hsu YC, Lin LT. Smart meter monitoring and data mining techniques for predicting refrigeration system performance. Expert Syst Appl 2014; 41: 2144–56.

[8] Fan C, Xiao F, Yan CC, Liu CL, Li ZD, Wang JY. A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. Appl Energy 2019; 235: 1551-60.

[9] Geronazzo A, Brager G, Manu S. Making sense of building data: New analysis methods for understanding indoor climate. Build Environ 2018; 128: 260–71.

[10] Afroz Z, Urmee T, Shafiullah GM, Higgins G. Real-time prediction model for

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

indoor temperature in a commercial building. Appl Energy 2018; 231: 29-53.

[11] Chen YB, Tan HW. Short-term prediction of electric demand in building sector via hybrid support vector regression. Appl Energy 2017; 204: 1363–74.

[12] Rafe Biswas MA, Robinson MD, Fumo N. Prediction of residential building energy consumption: a neural network approach. Energy 2016; 117: 84–92.

[13] Dong B, Cao C, Lee SE. Applying support vector machines to predict building energy consumption in tropical region. Energy Build 2005; 37: 545–53.

[14] Wang ZY, Srinivasan RS. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. Renew Sustain Energy Rev 2017; 75: 796-808.

[15] Wang ZY, Wang YR, Srinivasan RS. A novel ensemble learning approach to support building energy use prediction. Energy Build 2018; 159: 109–22.

[16] Fan C, Sun YJ, Zhao Y, Song MJ, Wang JY. Deep learning-based feature engineering methods for improved building energy prediction. Appl Energy 2019; 240: 35-45.

[17] Rahman A, Srikumar V, Smith AD. Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. Appl Energy 2018; 212: 372–85.

[18] Tian CL, Li CD, Zhang GQ, Lv YS. Data driven parallel prediction of building energy consumption using generative adversarial nets. Energy Build 2019; 186: 230-43.

[19] Gao DC, Wang SW, Shan K, Yan CC. A system-level fault detection and diagnosis method for low delta-T syndrome in the complex HVAC systems. Appl Energy 2016; 164: 1028–38.

[20] Thangavelu SR, Myat A, Khambadkone A. Energy optimization methodology for multi-chiller plant in commercial buildings. Energy 2017; 123: 64–76.

[21] Miller C, Nagy Z, Schlueter A. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings.

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

Renew Sustain Energy Rev 2018; 81: 1365–77.

[22] Fan C, Xiao F, Li ZD, Wang JY. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: a review. Energy Build 2018; 159: 296–308.

[23] Li MF, Ju YL. The analysis of the operating performance of a chiller system based on hierarchical cluster method. Energy Build 2017; 138: 695-703.

[24] Deb C, Lee SE. Determining key variables influencing energy consumption in office buildings through cluster analysis of pre- and post-retrofit building data. Energy Build 2018; 159: 228-45.

[25] Huang P, Sun Y. A clustering-based grouping method of nearly zero energy buildings for performance improvements. Appl Energy 2019; 235: 43-55.

[26] Yu Z, Haghighat F, Fung BCM, Zhou L. A novel methodology for knowledge discovery through mining associations between building operational data. Energy Build 2012; 47: 430-40.

[27] Fan C, Xiao F, Madsen H, Wang D. Temporal knowledge discovery in big BAS data for building energy management. Energy Build 2015; 109: 75-89.

[28] Xiao F, Fan C. Data mining in building automation system for improving building operational performance. Energy Build 2014; 75: 109-18.

[29] Fan C, Xiao F, Yan CC. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Automat Constr 2015; 50: 81-90.

[30] Funde NA, Dhabu MM, Paramasivam A, Deshpande PS. Motif-based association rule mining and clustering technique for determining energy usage patterns for smart meter data. Sustain Cities Soc 2019; 46: 101415.

[31] Capozzoli A, Piscitelli MS, Brandi S, Grassi D. Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings. Energy 2018; 157: 336-52.

[32] Fan C, Sun YJ, Shan K, Xiao F, Wang JY. Discovering gradual patterns in building

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

operations for improving building energy efficiency. Appl Energy 2018; 224: 116-23.

[33] Cook DJ, Holder LB. Graph-based data mining. IEEE Intell Syst App 2000; 15: 32-41.

[34] Nettleton DF. Data mining of social networks represented as graphs. Comput Sci Rev 2013; 7: 1-34.

[35] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. J Amer Soc Inform Sci Tech 2007; 58: 1019–31.

[36] Samatova NF, Hendrix W, Jenkins J, Padmanabhan K., Chakraborty A. Practical graph mining with R. 1st ed. Chapman & Hall/CRC; 2013.

[37] Aggarwal CC, Wang HX. Managing and mining graph data. Advances in Database Systems, Springer, 2010.

[38] Cook DJ, Holder LB. Mining graph data. 1st ed. New Jersey: Wiley; 2006.

[39] Jiang CT, Coenen F, Zito M. A survey of frequent subgraph mining algorithms. Knowl Eng Rev 2004; 0: 1-31.

[40] Ayed R, Hacid MS, Haque R, Jemai A. A list of FSM algorithms and available implementations in centralized graph transaction databases. Technical Report in The CAIR Project, 2016.

[41] Cook DJ, Holder LB. Substructure discovery using minimum description length and background knowledge. J Artif Intell Res 1994; 1: 231-55.

[42] Kuramochi M, Karypis G. GREW-A scalable frequent subgraph discovery algorithm. Proceedings of the 4th IEEE International Conference on Data Mining, 2004, 439-442.

[43] Borgelt C, Berthold M. Mining molecular fragments: Finding relevant substructures of molecules. Proceedings of International Conference on Data Mining, 2002, 211-218.

[44] Yan X, Han JW. gSpan: Graph-based substructure pattern mining. Proceedings of International Conference on Data Mining, 2002, 721-724.

[45] Huan J, Wang W, Prins J. Efficient mining of frequent subgraph in the presence

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

of isomorphism. Proceedings of the 2003 International Conference on Data Mining, 2003, 549-552.

[46] Nijssen S, Kok JN. A quickstart in frequent structure mining can make a difference. Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, 647-652.

[47] Worlein M, Meinl T, Fisher I, Philippsen M. A quantitative comparison of the subgraph miners MoFa, gSpan, FFSM and Gaston. Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2005, 392-404.

[48] Yan XF, Han JW. CloseGraph: Mining closed frequent graph patterns. The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August, 2003.

[49] Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. J Comput Graph Stat 2006; 15: 651-74.

[50] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, ISBN 3-900051-07-0; 2008. URL<http://www.R-project.org >.

[51] Csardi G, Nepusz. The igraph software package for complex network research. Inter Journal Compex Systems 2006; 1695.

[52] Pedersen TL. ggraph: An implementation of grammar of graphics for graphs and networks. The Comprehensive R Archive Network, 2018.

[53] Phillippsen M. ParSeMis – The parallel and sequential mining suite. Accessed on Jan 10, 2019, https://www2.cs.fau.de/En/resaerch/zold/ParSeMiS/Index.html.

✧ The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.