

Singapore Management University Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

12-2017

BTCl: A new framework for identifying congestion cascades using bus trajectory data

Meng-Fen CHIANG

Ee peng LIM


Singapore Management University, eplim@smu.edu.sg

Wang-Chien LEE

Agus Trisnajaya KWEE

DOI: <https://doi.org/10.1109/BigData.2017.8258039>

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the [Databases and Information Systems Commons](#), and the [Theory and Algorithms Commons](#)

Citation

CHIANG, Meng-Fen; LIM, Ee peng; LEE, Wang-Chien; and KWEE, Agus Trisnajaya. BTCl: A new framework for identifying congestion cascades using bus trajectory data. (2017). *Proceedings of 2017 IEEE International Conference on Big Data*. Research Collection School Of Information Systems.

Available at: https://ink.library.smu.edu.sg/sis_research/3971

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

BTCl: a New Framework for Identifying Congestion Cascades Using Bus Trajectory Data

Meng-Fen Chiang*, Ee-Peng Lim*, Wang-Chien Lee[†], and Agus Trisnajaya Kwee*

*Living Analytics Research Centre, Singapore Management University, Singapore

[†]Department of Computer Science and Engineering, The Pennsylvania State University, PA, USA

{mfchiang, eplim, aguskwee}@smu.edu.sg, wlee@cse.psu.edu

Abstract—The knowledge of traffic health status is essential to the general public and urban traffic management. To identify *congestion cascades*, an important phenomenon of traffic health, we propose a *Bus Trajectory based Congestion Identification (BTCl)* framework that explores the anomalous traffic health status and structure properties of congestion cascades using bus trajectory data. BTCl consists of two main steps, *congested segment extraction* and *congestion cascades identification*. The former constructs path speed models from historical vehicle transitions and design a non-parametric Kernel Density Estimation (KDE) function to derive a measure of *congestion score*. The latter aggregates congested segments (i.e., those with high congestion scores) into traffic congestion cascades by *unifying* both attribute coherence and spatio-temporal closeness of congested segments within a cascade. Extensive evaluations on 11.8 million bus trajectory data show that (1) BTCl can effectively identify congestion cascades, (2) the proposed congestion score is effective in extracting congested segments, (3) the proposed unified approach significantly outperforms alternative approaches in terms of extended precision, and (4) the identified congestion cascades are realistic, matching well with the traffic news and highly correlated with vehicle speed bands.

I. INTRODUCTION

Like human health, it is very important to know about the *traffic health* of roads in a city. From the general public point of view, knowledge of the traffic status allows road users to anticipate and avoid transportation delays. On the other hand, for urban traffic management, monitoring the traffic health is essential for transportation operators to optimize traffic flow. *Traffic congestion* is a phenomenon of traffic health that pretty much everyone in a crowded city has experienced with. Not only small disturbances in heavy traffic can be amplified into a traffic congestion, mis-timed traffic lights at junctions, car accidents, or road works can all cause congestions¹.

To detect congestions, a common practice today is via vehicle sensing, e.g., Pneumatic road tube counting, which requires expensive installation at roads or junctions and usually is not effective on high-volume, multi-lane highways, where congestions often take places. Moreover, the reporting of accidents or traffic congestions heavily relies on expert judgment and decision. Due to subjective and imprecise human judgement, traffic congestion events may not be consistently and accurately detected and reported. Furthermore, existing traffic health status is usually presented in forms of speed

indicators, e.g., Waze² and Google Map provide live map of traffic health status reports, which color low speed road segments in red. These speed indicators, however do not address the formation and lifecycle of some congestion events. Thus, finding an effective solution to *identify* traffic congestion events, and potentially leading to explanation of their happenings, remains an important open problem.

In this work, we propose to explore explanation data to identify traffic congestion events, which serves as the first step to better understand the traffic congestion event in order to provide a timely traffic health monitoring and reporting. To illustrate this idea, we utilize thousands of buses equipped with GPS systems to generate a tremendous amount of bus trajectory data. Treating these buses as mobile sensors, the collected bus trajectory data provides the “pulses” of the traffic health and has a potential to be exploited for detection of traffic congestions, which is a nontrivial problem. Notice that a traffic congestion does not only exist in one single location point. It usually grows into a region and continues to expand over time before it eventually shrinks geographically and disappears. We term this lifecycle of congestion expansion and shrinking as *congestion cascade*.

Example I.1. (Congestion Cascade) Figure 1 illustrates a real example of traffic congestion on 2016-05-25. An accident happened at 16:01 near location point A. After then, the traffic piled up from A towards C and progressively resulted in congestions over time. The formation of the congestion from A towards B was reported at 16:16, followed by another report from A to C at 16:34. While this congestion event may be manually identified, it would be much more cost effective to utilize sampled vehicle trajectory data to uncover the above and many other similar events. For illustration, suppose we have two overlapping road segments as SEG0 and SEG1 covering [F,E,D,C] and [E,D,C,B], respectively. On 25 May 2016, Figure 1(b) reveals a speed slowdown on SEG0 after 16:00 (median speed is less than 5 km/hr) with significant speed drop compared with its historical norm shown in Figure 1(c) (median speed is around 10 km/hr). Figure 1(d) also reveals a speed slowdown on SEG1 after 16:15 (after some minutes after), with significant speed drop compared with its historical norm in Figure 1(e). These observations, collectively indicating the formation of a traffic congestion, motivating us to exploit

¹https://en.wikipedia.org/wiki/Traffic_congestion

²<https://www.waze.com/livemap>



(a) Reported Traffic News

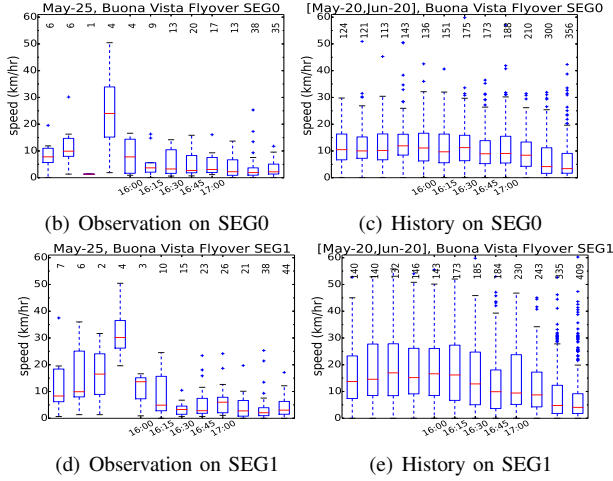


Fig. 1: Speed Drop of Traffic Flow from SEG0 to SEG1

vehicle trajectories to identify traffic congestion cascade.

Congestion cascade is essentially a spatio-temporal phenomena of traffic health. We therefore formulate the research problem of identifying congestion cascades as follows.

Problem I.1. (Congestion Cascade Identification)

Given: a database of historical vehicle trajectories H and a target set of vehicle trajectories D collected over a time period of interest T .

Identify: congestion cascades within T by considering potential factors such as direction of congestion influence, spatio-temporal closeness, and congestion level compared with the relevant historical trends.

Assuming no prior knowledge of incidents, the congestion cascade identification problem is very challenging, as we seek to identify the lifecycle and spatial scope of congestion cascades instead of only highlighting the evidence of individual congestions (e.g., slow and abnormal traffic flow of road segments as Waze does). To achieve our goal, we characterize traffic congestion cascades by taking into account the anomalous traffic health status and structure properties of congestion cascades.

To achieve our goal, we propose a novel *Bus Trajectory based Congestion Identification (BTCI)* framework that consists of two major components: (1) congested segment extraction, and (2) congestion cascade clustering. In BTCI, we first leverage spatial and temporal dependencies embedded in historical vehicle trajectory data H to statistically capture the general (normal) traffic health on road segments at all time windows (called *spatio-temporal segments* or simply *segments*). Compared against the norm of traffic, we propose a novel statistics-based method to quantify a *congestion score* for each segment during the time period of interest T using the targeted

set of vehicle trajectories D . Next, we propose a clustering algorithm to aggregate *congested segments* (i.e., those with high congestion scores) into traffic congestion cascades by considering the structure properties and abnormal traffic health status of potential congestion cascades.

Contributions. This paper addresses the challenges in congestion cascades identification and makes the following contributions:

Approach. We exploit vehicle trajectory data to cost-effectively identify congestion cascades.

Framework. We propose the BTCI framework to identify congestion cascades, which include innovative algorithms for congested segment extraction and congestion cascade construction.

Concepts. We propose novel concepts that best captures congestion cascades, including structure properties (i.e., direction of congestion influence and spatio-temporal closeness) and anomalous traffic health (i.e., congestion scores).

Unified Generative Model. We adopt a generative model to learn cluster memberships for congested segments based on both attribute similarities and structure closeness.

Experimentation with Real Data. Experiments on bus trajectory data show that the BTCI framework is able to identify congestion cascades with coherence in space, time, direction of traffic flow, and congestion scores among engaged segments.

The remainder of this paper is organized as follows. Section II presents our framework. Section III introduces our method for congestion score estimation. Section IV presents the proposed unified approach for congestion cascade identification. Section V evaluates our framework using real-world dataset. Section VI reviews the related work. Section VII concludes this study.

II. PRELIMINARY

In this section, we formally define terminologies and give an overview of our proposed framework. Table I summarizes the notations to be used.

A. Definitions

Definition II.1. (Road Network) A road network is a directed graph $RN=(V,E)$, where each node $v \in V$ represents an end point of some road segment, and each directed edge $e \in E$ is a stretch of road segment connecting point e_s to next point e_d .

Definition II.2. (Path) A path is a sequence of connected edges, i.e., $r: (e_1, e_2, \dots, e_n)$, where $e_{i+1}.s = e_i.d$ ($1 \leq i < n$). The length of path r is measured by its number of nodes, i.e., $|r| = n+1$. A path r is essentially a sequence of road segments where traffic flows from $e_1.s$ to $e_n.d$.

Definition II.3. (Vehicle Transition) Given a path r and two adjacent timestamps, t and $t+1$, a vehicle transition on path r is denoted as $(id, loc_r^t, loc_r^{t+1})$, where id is the vehicle ID; (loc_r^t, loc_r^{t+1}) are the locations of the vehicle on r at timestamps t and $t+1$.

Definition II.4. (Speed) The speed of a vehicle transition $(id, loc_r^t, loc_r^{t+1})$ is estimated by $\Delta(loc_r^t, loc_r^{t+1})/\Delta(t, t+1)$,

where $\Delta(\text{loc}_r^t, \text{loc}_r^{t+1})$ is the spatial distance and $\Delta(t, t+1)$ is the time difference.

Obviously the behavior of vehicle transitions on bustling road segments in morning peak hours is different from that on quiet road segments in late evenings. Capturing varied vehicle transition behaviors across different time periods and road segments is essential for precise speed modeling and congestion cascade identification. Hence, we consider vehicle transitions within a spatio-temporal unit of traffic, called *spatio-temporal segment* or *segment* in short, as follows.

Definition II.5. (Segment) A segment $(r, [t^*, t^* + \delta t])$ is a spatio-temporal unit of interest, where r is a path, and $[t^*, t^* + \delta t]$ is a time window from a time-of-the-day t^* to another time-of-the-day $t^* + \delta t$.

Note that we use t to denote a time point, δt to denote a time interval, and t^* to denote a time-of-the-day.

Example II.1. Suppose the time window is 15 mins. Let the time window slide every minute. Then, there are 1440 different time windows $T = \{[00:00,00:15], \dots, [23:59,00:14]\}$ in a day. Suppose we have a set of possible road paths $R = \{(e_1, e_2, e_3), (e_2, e_3, e_4), \dots\}$ on the road network. In total, we obtain $|T| \times |R|$ segments, $S = \{S_1, \dots, S_{|T| \times |R|}\}$, where each segment $S_i = (r_i, [t_i^*, t_i^* + \delta t])$ is associated with a road path $r_i \in R$ and a time window $[t_i^*, t_i^* + \delta t] \in T$. We say two segments are spatio-temporally overlapped if both their paths and time windows are partially overlapped, e.g., (e_1, e_2, e_3) spatially overlaps (e_2, e_3, e_4) and $[10:00,10:15]$ temporally overlaps $[10:01,10:16)$. Note that the path length and time window can be empirically determined as long enough to observe vehicle movements within the path and a time window.

Definition II.6. (Congested Segment) A congested segment (also called *c-segment*) $S_i = (r_i, [t_i^*, t_i^* + \delta t])$ is a segment within a target day with a high congestion score C_{S_i} where C_{S_i} is greater than a given congestion threshold ϵ .

As observed in Example I.1, a congestion may be contagious spatially and temporally over nearby segments. Moreover, the traffics of segments in a congestion cascade flow in the same direction. We define a congestion cascade as follows.

Definition II.7. (Congestion Cascade) A congestion cascade consists of a set of congested segments with spatio-temporal closeness and coherent traffic flow direction.

B. An Overview of the BTCI Framework

The main idea of the proposed BTCI framework is to extract congested segments, by comparing against the statistical norm of traffics, derived from the historical vehicle transitions database H (which can be seen as a sequence of snapshots S^1, \dots, S^t of vehicle locations over the road network) from a target set of vehicle transitions D during a target day of interest T (which consists of a sequence of snapshots $S^{t'+1}, \dots, S^{t'+m}$ where $t' \geq t$). Figure 2 depicts the proposed framework that consists of two components: (1) congested segment extraction, and (2) congestion cascade clustering.

TABLE I: Summary of Notation

Sym.	Definition
t	a time point
$[t^*, t^* + \delta t]$	a time window starts at t^* and ends at $t^* + \delta t$
r_i	a path e_{i1}, \dots, e_{ik}
S^t	a trajectory snapshot at time t
$S_i = (r_i, [t_i^*, t_i^* + \delta t])$	a segment on r_i during $[t_i^*, t_i^* + \delta t]$
C_{S_i}	congestion score of segment S_i
$CG = \{S_1, \dots, S_q\}$	a congestion cascade
H	vehicle transitions database
D	vehicle transitions of target day

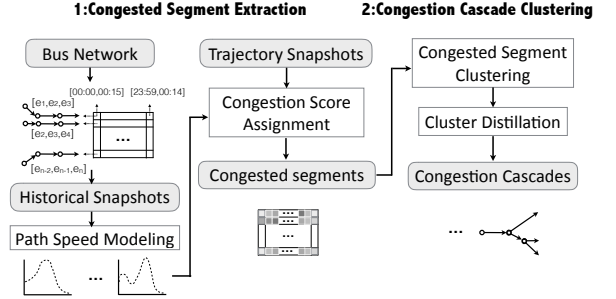


Fig. 2: Overview of the BTCI Framework

Congested Segment Extraction. This component aims to assign a congestion score to each segment in D . First, we derive path speed model that captures the speed norm of vehicles for each segment from H . We argue that a congestion is a slow-speed anomaly different from the usual speed pattern. For each segment $S_i = (r_i, [t_i^*, t_i^* + \delta t])$, we adopt a *non-parametric* approach to statistically derive a *path speed model*, which does not assume fixed number of clusters in the model and thus can better accommodate the complexity of the data. Second, we consider the spatio-temporal dependency of speed patterns and adaptively determine a congestion score for each segment.

Congestion Cascade Clustering. As the traffic flow of nearby segments may affect each other, we assume that a congestion cascade constitutes a group of spatially and temporally clustered congested segments. Thus, this component aggregates relevant cascades through *congested segment clustering* such that each resultant congestion cascade shows coherent structure properties and anomalous traffic health status.

III. CONGESTED SEGMENT EXTRACTION

Problem Analysis. As discussed, our idea is to establish the statistical norm of traffic speed based on historical data H and then use an effective scoring function to measure the slow-speed anomaly of a segment based on the deviation of observed speeds D on the target day from its norm with the same (road path, time window) pair. The extraction of congested segments is carried out in two steps: (1) we pre-compute the path speed model for each segment S_i to assess the congestion score for the observed speed data of segment S_i , (2) we determine the speed threshold for segment S_i using the path speed model f_{S_i} of the same (road path, time window) pair and derive the congestion score based on speed observations of segment S_i . The larger the congestion score, the more likely the segment S_i is congested. The problem of congested segment extraction is defined as follows.

Problem III.1. (Congested Segment Extraction) Given a set of

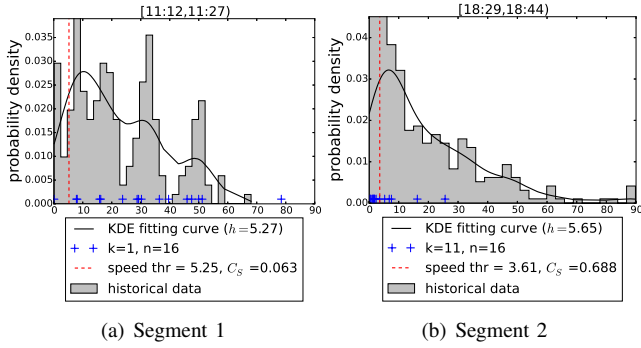


Fig. 3: Examples of Path Speed Modeling.

segments S_1, \dots, S_p , find a set of congested segments, denoted as *c-segments*, $CS = \{S_i | C_{S_i} \geq \gamma\}$ where C_{S_i} denotes the congestion score of segment S_i and γ denotes the score threshold for qualifying congested segments. γ is to be determined empirically.

A. Modeling of Path Speed

We aim to derive a speed model for each unit path r at a time window $[t^*, t^* + \delta t]$ from historical snapshots. We first extract all vehicle transitions on road r from the snapshots and derive speed data of vehicle transitions, denoted as $\{s_1, \dots, s_n\}$, where s_j is the speed of a vehicle transition on path unit r at the above time window.

Non-Parametric Approach: We propose to profile each segment $S_i = (r_i, [t_i^*, t_i^* + \delta t])$ using $\{s_1, \dots, s_n\}$ by Kernel Density Estimation (KDE). Rather than prescribing some known distribution (e.g., Gaussian with a single mode), KDE [1] uses the sum of kernel functions centered at data points to estimate the density. Typically, an isotropic Gaussian kernel is chosen for each training point, with a single shared hyperparameter. The estimate of the probability density at each point in the data space relies on data points lying within its neighborhood, specified by the *bandwidth* of the kernel. The bandwidth of isotropic Gaussian kernel h determines how widely it expands from the data point. A kernel density estimator is then trained by determining the variance of the kernels, which controls the smoothness of the overall distribution. Allowing adaptive kernel bandwidth over the data space of vehicle speeds, our speed estimators can capture traffic conditions that vary among segments. To avoid overfitting or underfitting, the kernel bandwidth \hat{h}_{S_i} is determined by maximizing the cross-validated likelihood for historical speed data $\{s_1, \dots, s_n\}$ for each segment S_i .

B. Adaptive Congestion Score

Based on the path speed model obtained for a segment S_i (with path r_i in $[t_i^*, t_i^* + \delta t]$), we assign a *congestion score* to the corresponding segment in target transition set \mathcal{D} , such that the larger the congestion score is, the more likely the segment S_i is congested.

To derive the congestion score from speed observations in segment S_i , we consider the probability of a new observation to belong to slow-speed anomaly. Namely, we use the lower tail region (e.g., the slowest 10% of the speed distribution) to represent the area of anomalous slow speeds. As such, a

speed observation is considered as an evidence of congestion if the value falls within the lower tail region. Suppose S_i has k_i speed observations falling within the lower tail region, the congestion score of S_i is then defined to be $C_{S_i} = \frac{k_i}{n_i}$, i.e., the proportion of anomalous slow speed observations.

Example III.1. Figure 3 shows two examples of path speed models using KDE based on one-month historical speed data for two segments. Figure 3(a) depicts the model for the segment during [11:12,11:27) on one path. The grey histogram is the distribution of historical speed data on the segment, and the curve is the KDE after fitting historical data with an adaptive kernel ($\hat{h}_{S_i} = 5.27$). The speed observations are depicted in blue + marks. We observe three modes in [11:12,11:27). Take a lower tail threshold $c = 10\%$ as an example. We obtain a speed threshold at 5.25 km/hr (indicated by red vertical line). The congestion score of this segment is $C_s = 0.063$, as 1 out of 16 speed observations in the segment falls below the speed threshold. Figure 3(b) depicts the model for the same path during another time window [18:29,18:44). As shown the historical speed skewed to the slowest mode, we have a lower speed threshold at 3.61 km/hr. As there are 11 out of 16 speed observations falling below the threshold, this segment has a higher congestion score $C_s = 0.688$.

IV. CONGESTION CASCADE IDENTIFICATION

Congestions are contagious as a congested path may cause the neighboring paths to be congested. The phenomena of a congestion cascade involves several neighboring paths congested around the same time (or for a period of time), i.e., each road path involved in the cascade has one or more *c-segments* detected. Additionally the traffic flows of involved *c-segments* follow the same direction. Accordingly, we refine the definition of congestion cascade as follows.

Definition IV.1. (Congestion Cascade) A congestion cascade is a weighted graph of *c-segments* $CG = \{S_1, \dots, S_m\}$, where a node represents a *c-segment* S_i and an edge (S_i, S_j) reveals that S_j is influenced (infected) by S_i . The influence relation between two desired *c-segments* suggests the following two properties of a congestion cascade: (1) coherent traffic flow direction, and (2) spatio-temporal closeness among the two *c-segments*.

For example, consider SEG0 and SEG1 in Figure 1. We observe congestion from 16:00 to 18:00, and we obtain two *c-segments*, (SEG0, [16:00,18:00)) and (SEG1, [16:00,18:00)) respectively. A congestion cascade consisting of the two *c-segments* can be identified based on the spatial and temporal closeness of the *c-segments*.

Problem IV.1. (Congestion Cascade Identification)

Given: a set of *c-segments* $CS = \{S_i | C_{S_i} \geq \gamma\}$.

Find: a set of congestion cascades $\{CG_1, \dots, CG_q\}$ based on Definition IV.1.

In the following, we propose three approaches to identify congestion cascades: (1) connectivity-based, and (2) attribute-

based, and (3) unified generative model approaches.

A. Connectivity-Based Approach

To find congestion cascades that satisfy the two properties in Definition II.7, an idea is to encode *flow direction* and *spatio-temporal closeness* between a pair of c-segments as a relationship, and thus form connected graphs of c-segments. Naturally, each of those connected graphs represents a congestion cascade by satisfying the requirements. To encode both flow direction and closeness in spatio-temporal in the connection between two c-segments, we define *segment overlap* as follows:

Definition IV.2. (Segment Overlap) A pair of c-segments $S_i = (r_i, [t_i^*, t_i^* + \delta t])$ and $S_j = (r_j, [t_j^*, t_j^* + \delta t])$ are *connected*, denoted as $e(S_i, S_j)$, if (1) r_j and r_i overlaps spatially, and (2) $[t_i^*, t_i^* + \delta t]$ and $[t_j^*, t_j^* + \delta t]$ overlaps temporally.

As such, the overlaps among c-segments can be represented as a set of connected components $\{CG_1, \dots, CG_q\}$, where each connected component $CG_k = \{S_{k1}, \dots, S_{kl}\}$ consists of nodes for c-segments S_{ki} , and edges for each pair of overlapped c-segments among S_{k1}, \dots, S_{kl} . We perform breadth-first search (BFS) to collect connected components CG_1, \dots, CG_q . The size of a congestion cascades $CG_k = \{S_{k1}, \dots, S_{kl}\}$ is l . The *spatial scope* of CG_k is defined as the union of paths in c-segments $\cup_{i=1}^l r_{ki}$. The *temporal scope* of CG_k is defined as the interval $[\min(\{t_{ki}^*\}), \max(\{t_{ki}^*\}) + \delta t]$, $1 \leq i \leq l$.

Example IV.1. Consider again the example in Figure 1. Note that the upstream segment [E,D,C,B] at time [16:15,16:30] and the downstream segment [F,E,D,C] at time [16:15,16:30] share the same traffic flow direction and are spatio-temporally close. Thus, the connectivity-based approach considers them as belonging to the same congestion cascade.

B. Attribute-based Approach

While the connectivity-based approach considers both temporal and spatial connectivities among congested segments of a congestion cascade, it does not consider coherence of attribute values associated with congested segments, e.g., segment's congestion score, average speed of bus observations, and so on. Therefore, we introduce the attribute-based approach to generate cascades with attribute coherence.

A few clustering methods can be adopted to group congested segments into congestion cascades (e.g., k-means, K-Medoids, spectral clustering, DENCLUE [2]). A challenge arising in our problem setting is that different attribute domain has its own numeric scale and data distribution. As DENCLUE can better capture major peaks of data according to individual distribution of attribute domains, we thus adopt it to derive attribute-based clusters. In our problem setting, each c-segment is represented as an attribute vector consisting of the start and end location coordinates of path, start and end timestamps, traffic flow direction, and congestion score. These attributes can be categorized into four aspects: spatial, temporal, traffic flow direction, and congestion score aspects.

DENCLUE utilizes two parameters to derive local maxima of the density function as clusters. The first parameter determines

the influence of a point in its neighborhood and the second one describes whether a density-attractor is significant. Please refer to [2] for details on parameter settings. As DENCLUE is not controlled by number of clusters, we first find local maxima of the density function as centroids in each aspect. Then, we enumerate all combinations of centroids in each aspect. For example, suppose we have two spatial centroids and two temporal centroids. Then we obtain in total four centroids (2×2) in the spatio-temporal space. Typical temporal centroids, morning and evening peak hours, are observed from the start and end time of c-segments. Lastly, we randomly choose K initial centroids and assign memberships using k-means algorithm to K clusters.

C. Unified Generative Model

Neither connectivity-based nor attribute-based approaches consider both spatio-temporal connectivities and attribute coherence together. The connectivity-based approach also does not consider the *weights* of connections that indicate the level of spatial and temporal overlaps between congested segments. In addition, some c-segments in reality may not be involved in major congestion cascades. Including such c-segments in congestion cascades only introduces noises and results in incoherences. To address these pitfalls, we introduce a unified congestion cascade identification approach designed to derive congestion cascades that *exhibit similar attributes' values as well as strong spatial and temporal connectivity*.

The unified approach represents the input set of congested segments as a 4-tuple graph $G = (CS, E, A, W)$ where CS and E denote the set of congested segments and edges of connected segment pairs defined by spatial and temporal overlaps. A denotes a set of segment attributes $\{a_1, \dots, a_j\}$ associated with nodes in CS where $a_k(S_i)$ denotes the a_k attribute value of c-segment S_i . W , expressed by a connection weight function $w(S_i, S_j)$, denotes the set of weights corresponding to edges in E . We aim to take into account both attribute coherence and connection strength into clustering with the following properties: (1) c-segments with similar attribute values have similar clustering membership, and (2) strongly adjacent c-segments have similar clustering membership.

To achieve our goal, we adopt GenClus, a general soft clustering approach developed for attributed heterogeneous networks [3]. In our problem setting, the c-segment graph consists of vertices with a number of attributes, including *start and end location coordinates of path, start and end timestamps, traffic flow direction, and congestion score* and weighted edges representing direction-aware spatio-temporal closeness. The core of the unified approach is a probabilistic generative model that assumes *attributes values and connectivity strength are generated corresponding to a clustering*. The generative process is as follows. Consider K clusters, each of which represents a congestion cascade. Each cascade has j latent factors (also called *cluster attribute parameters*), denoted by β that influence (i.e., probabilistically generate) the attributes of c-segments. Here β is modeled by a combination of j univariate Gaussian distributions over K clusters. Thus, for

each c-segment S_i , we sample a cluster k from K clusters according to a mixture weight estimated by the similarity in both attribute and clustering membership in each cluster. Given cluster k and its cluster attribute parameters β_k , we then sample each attribute for S_i .

Given the spatio-temporal connectivity W , and the cluster attribute parameters β , the likelihood of observing all attribute values Λ and a clustering Θ can be formally modelled as follows.

$$p(\Lambda, \Theta|W, \beta) = \prod_{\lambda \in \Lambda} p(\lambda|\Theta, \beta)p(\Theta|W) \quad (1)$$

where $p(\lambda|\Theta, \beta)$ is the generative probability of observing an attribute value $\lambda \in \Lambda$, the given Θ and β , and $p(\Theta|W)$ is the probability of Θ given graph connectivity structure W . The goal is to find the best clustering that maximizes the likelihood in Eq. (1), which entails finding the best parameters β and Θ . The generation of attributes and connectivity are considered separately.

Modelling Attribute Generation. In BTCl, we consider the following eight attribute domains: (1) congestion score, (2) latitude of start point, (3) longitude of start point, (4) latitude of end point, (5) longitude of end point, (6) traffic flow direction in degree, (7) start time of a segment, and (8) end time of a segment. We assume each attribute domain follows a mixture of K Gaussian distributions $\beta_k=(\beta_{k1}, \dots, \beta_{k8})$, $1 \leq k \leq K$, where each attribute domain a_j in a cluster k follows a Gaussian distribution with parameters $\beta_j=(\mu_{kj}, \sigma_{kj}^2)$, i.e., $\lambda_{kj} \sim \mathcal{N}(\mu_{kj}, \sigma_{kj}^2)$, $1 \leq j \leq 8$. μ_{kj} and σ_{kj} are mean and standard deviation of Gaussian distribution for attribute domain a_j in cluster k . $\theta_{v,k}$ is the membership of c-segment v in cluster k . The probability density for the observed attribute values of all c-segments in a given clustering Θ is then:

$$p(\Lambda|\Theta, \beta) = \prod_{j=1}^8 \prod_{v \in CS} \sum_{k=1}^K \theta_{v,k} \frac{1}{\sqrt{2\pi\sigma_{kj}^2}} e^{\left(-\frac{(\lambda_v - \mu_{kj})^2}{2\sigma_{kj}^2}\right)} \quad (2)$$

where CS is the set of c-segments, Θ is a clustering assignment, β is the cluster attribute parameters, and attributes are assumed to be independent from each other.

Modelling Connectivity Generation. Based on GenClus, the connectivity strength of the c-segment attributed graph is also generated from the clustering. The model assigns to every c-segment v a distribution of cascade memberships over K cascades, $\theta_{v,k}$, $1 \leq k \leq K$. The idea is that if an edge weight $w(v_i, v_j)$ of given two c-segments v_i and v_j is greater and their cluster memberships θ_i and θ_j are similar, they are considered as being more consistent with each other and thus more likely to belong to the same cluster. To measure the consistency of a clustering Θ , GenClus utilizes a cross-entropy-based probability density function for every adjacent pair θ_i, θ_j as follows.

$$f(\theta_i, \theta_j, w(v_i, v_j)) = w(v_i, v_j) \sum_{k=1}^K \theta_{j,k} \log \theta_{i,k} \quad (3)$$

where $f(\theta_i, \theta_j, e)$ yields a greater value if the clustering memberships of v_i and v_j are more similar and they are connected with a greater adjacent strength $w(v_i, v_j)$. GenClus utilizes a log-linear model to derive the probability of Θ given the edge weights W as follows.

$$p(\Theta|W) = \exp \left(\sum_{w(v_i, v_j) \in W} f(\theta_i, \theta_j, w(v_i, v_j)) \right) \quad (4)$$

Combination. Eq. (1) can be easily obtained by combining Eq. (2) and Eq. (4).

In the following, we discuss several issues in the process of model parameter learning which are carried out by BTCl to identify congestion cascades.

Cluster Initialization. Good initial clusters are essential for final clustering quality. There are a number of studies on selecting good centroids [2][4]. We follow the idea of density-based clustering in Section IV-B to obtain initial K clusters. **Cluster Optimization.** The goal in cluster optimization is to utilize both connectivity structure and attribute information to derive the best clustering for the graph G . GenClus is an iterative algorithm that alternatively optimizes clustering to maximize the objective function in Eq. (1). We adopt EM algorithm to optimize clusters iteratively. Please refer to GenClus [3] for details.

Distilling Congestion Cascades. Major clusters that are highly coherent in attribute values and spatio-temporal closeness are promising candidates of congestion cascades. However, some c-segments in reality may be disengaged in some congestion cascades, i.e., some c-segments may not be part of any congestion cascades. These cases represent noises and introduce incoherences in the resultant clustering. To distill truly interesting congestion cascades, we introduce two post-processes to tackle (1) *weak c-segments* and (2) *incoherent c-segments* as follows.

Weak C-Segment Filtering. Once we obtain cluster membership for each c-segment $v \in CS$, we assign v exclusively to the cluster with the strongest membership, i.e., $\kappa = \arg\max_k p(z_v = k|\Theta, \beta)$. Some c-segments unfortunately have very small likelihood to its assigned cluster κ and are known as weak c-segments. Weak c-segments may represent noises to congestion cascades. Thus, we introduce a *membership threshold* (ϵ) to eliminate weak c-segments from hard clustering result (i.e., $\epsilon \leq p(z_v = \kappa|\Theta, \beta)$).

Incoherent C-Segment Filtering. Some c-segments are engaged in its belonging clusters due to partial coherence in a subset of attribute values and/or weak spatio-temporal adjacency. We refer these c-segments as *incoherent c-segments*. The dissimilar attribute values of incoherent c-segments present incoherence of corresponding attribute domains to belonging cluster. For example, noisy c-segments, with low generative probability in temporal space but high generative probabilities in other spaces, contribute temporal incoherence to belonging cluster. To distill clusters with higher overall attribute coherence, we introduce a *coherence threshold* (φ) to recursively perform soft-clustering on incoherent clusters until each sub-cluster satisfies specified φ (i.e., $\sigma_k^i \leq \varphi, \forall 1 \leq i \leq 8$).

Algorithm 1 summarizes the unified approach for detecting congestion cascades with integration of proposed distilling strategies. Line 1 is the cluster initialization. Line2-5 illustrate the cluster optimization process. Lines 6-13 depict the process to distill congestion cascades.

Algorithm 1: Congestion Cascade Identification

Input: $G=(CS, E, A, W)$: c-segment attributed graph, K : cluster number, ϵ : membership threshold, φ : coherence threshold;

Output: CG : the set of congestion cascades, β : attribute component parameters;

```

/* step 1: cluster initialization */
1  $\Theta, \beta \leftarrow \text{Initialization}(G)$ ;
/* step 2: cluster optimization */
2 repeat
3    $\text{updateAssignments}(\Theta^{t-1}, \beta^{t-1}), \forall v \in CS$ ;
4    $\text{updateParameters}(p(z_v^t | \Theta^{t-1}, \beta^{t-1})), \forall v \in CS$ ;
5 until maximum iteration;
/* step 3: distilling congestion cascades */
6  $\Theta_\epsilon \leftarrow \text{weakC-SegmentFiltering}(\Theta, \epsilon)$ ;
7  $\Theta_\epsilon^N \leftarrow \text{getNoisyCluster}(\Theta_\epsilon)$ ;
8 for  $\theta_\epsilon^N \in \Theta_\epsilon^N$  do
9   repeat
10     $\theta_\epsilon^{N'} \leftarrow \text{soft-clustering}(\theta_\epsilon^N)$ ;
11     $\Theta_\epsilon^G \cup \text{getCoherentCluster}(\theta_\epsilon^{N'}, \varphi)$ ;
12     $\theta_\epsilon^N \leftarrow \text{getNoisyCluster}(\theta_\epsilon^{N'}, \varphi)$ ;
13   until  $\theta_\epsilon^N$  is empty;
14 return  $(\Theta_\epsilon - \Theta_\epsilon^N) \cup \Theta_\epsilon^G$ ;

```

Example IV.2. Figure 4 illustrates the resultant cascades of a c-segment attribute graph using three proposed approaches. In the graph, each node is a c-segment associated with one attribute value and each edge is weighted by the spatio-temporal closeness between two connected c-segments. In Figure 4(a), connectivity-based approach returns the entire graph as a cascade. In Figure 4(b), attribute-based approach identifies $\{S_1, S_2, S_3\}$ and $\{S_0, S_4, S_5\}$ as cascades by attribute coherence. $\{S_0, S_4, S_5\}$ is unsatisfactory as S_0 is spatio-temporally distant from S_4 and S_5 . Lastly, in Figure 4(c) unified approach returns $\{S_0, S_1, S_2\}$ and $\{S_3, S_4, S_5\}$ considering both spatio-temporal connectivity and attribute coherence in cascades identification.

V. PERFORMANCE EVALUATION

In this section, we evaluate the proposed BTCI framework using real datasets and report empirical findings in the following aspects: (1) evaluation on c-segment extraction, (2) evaluation on congestion cascade identification, (3) case studies, and (4) sensitivity tests on parameters.

A. Datasets and Settings

Dataset. We acquired a dataset consisting of **bus trajectory data**³ and the **bus stop network** of Singapore. There are two sets of bus trajectory data from different time periods: one from 2016-05-20 to 2016-06-20, and the other for the entire July 2016, known as the *May-June* dataset and *July* dataset, respectively. We extract the bus transitions from the snapshots of bus locations. The May-June and July datasets contain 25,611 and 24,973 snapshots, respectively. We then derive the speed observations from the snapshots for the May-June and July datasets.

In this paper, we focus on bus transitions on a highway in Singapore. We obtain 11.8 million bus transitions from all buses passing through the highway which has 18 bus

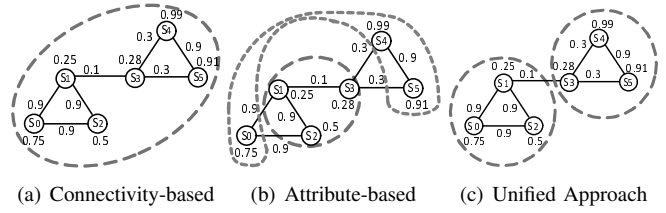


Fig. 4: Congestion Cascade Identification Approaches

stops. Accordingly, we derive all paths corresponding to road segments covering four bus stops (note that each path is overlapped with nearby paths with some bus stop difference). There are 82 such paths. We then adopt a 15-minute sliding time window with 1-minute shift over the the entire day to generate 1440 different time intervals. Hence, there is a total of 118,080 (82×1440) segments for a day. After discarding segments with less than ten speed observations, we obtain 87,052 segments from the May-June dataset. The speed observations from the May-June dataset are used for constructing the path speed models of segments, while that from the July dataset are for evaluating the BTCI framework, i.e., assessing the performance of c-segment extraction and congestion cascade identification.

B. Evaluation on C-Segment Extraction

The goal of this evaluation is to demonstrate that the proposed *congestion score* is a good indicator of congestions. As we do not have the complete ground-truth of traffic health status for all paths (and segments) on the examined highway, we resort to the use of an external source of **vehicle speed band data** which express the average speed of general vehicles on different road segments of the highway in 4 bands, $[0,20)$, $[20,40)$, $[40,60)$, and $[60, \infty)$, denoted by 1, 2, 3 and 4, respectively. For each road segment r' , the speed band data contains a time series of $(r', t_j, \text{speed_band}_j)$ tuples, where speed_band_j is the band value. Due to mismatch in spatial scope of the speed band data and our spatio-temporal segments, we compute the average speed band value for each segment $(r, [t^*, t^* + \delta t])$.

Congestion Score v.s. Vehicle Speed. While we do not have the complete ground truth of congestions for the highway under examination, we explore in our evaluation a common knowledge that *vehicles speeds in a congestions are slow*. In other words, an effective congestion indicator is supposed to be highly correlated to vehicle speeds in congestions. Thus, we measure the Pearson correlation coefficient between average vehicle speed band data and congestion scores of segments and plot the result in Figure 5, using data in five randomly selected days in July with known congestion incidents. For comparison, we also include the correlation of *average bus speed*, an alternative congestion indicator, with the speed bands. Our result shows a negative correlation between congestion score and average speed band value (i.e., the higher the congestion score, the lower the speed), while the average bus speed shows a positive correlation with the speed band value⁴. By varying the low tail threshold, we observe the strongest correlation 0.535 occurs when $c=0.15$. Generally speaking, the correlation

³We could not reveal the source due to non-disclosure agreement.

⁴For ease of comparison, we show the absolute value of correlations.

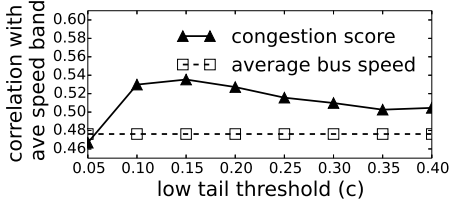


Fig. 5: Correlation with Average Speed Band

of the congestion score is stronger than that of average bus speed, which suggests that average bus speed may not be as good a congestion indicator as the proposed congestion score.

C. Evaluation on Congestion Cascade Clustering

We conduct experiments to evaluate the accuracy of identified congestion cascades against traffic news reported by the traffic authority⁵. Table II shows two reports of traffic news on 2016-07-20. To evaluate against the traffic news reports, we rank all identified congestion cascade by a ranking function $D \times wScr$ where D is the graph density of a cascade and $wScr = \frac{1}{|CG_i|} \sum_{S \in CG_i} p(z_S = \kappa | \Theta, \beta) \times C_S$, i.e., the average congestion score of c -segments in the cascade CG_i weighted by the membership to the cascade. Intuitively, a cascade with spatially and temporally concentrated (i.e., higher D) and severely congested (i.e., higher $wScr$) c -segments is ranked higher. We also introduce *extended precision* (EP) to measure the effectiveness of identifying and ranking congestion cascades to catch congestion events reported in traffic news.

Extended Precision (EP). Extended precision (EP) considers the coverage ratio of the spatial and temporal scopes of congestion reported by the traffic news. Let a cascade $CG_i = \{S_{i1}, \dots, S_{in}\}$ be associated with a spatial scope $R_i = \cup_{l=1}^n r_{il}$ and a temporal scope $T_i = [\min(\{t_{il}^*\}), \max(\{t_{il}^*\}) + \delta t]$, $1 \leq l \leq n$. We define the ratio of *spatial coverage* between a traffic news $E_j = (r_j^g, t_j^g)$ and a cascade CG_i as follows.

$$\mathbb{R}^S(R_i, r_j^g) = \frac{R_i \cap r_j^g}{R_i \cup r_j^g} \quad (5)$$

where $R_i \cap r_j^g$ and $R_i \cup r_j^g$ are the spatial intersection and union of CG_i and E_j , respectively. $\mathbb{R}^S(R_i, r_j^g)$ ranges within (0,1), which gives a higher ratio when the cascade CG_i greatly overlaps with the traffic news E_j by space. The *temporal coverage* $\mathbb{R}^T(T_i, t_j^g)$ between a ground-truth time-stamp E_j and an identified temporal scope CG_i is given as follows.

$$\mathbb{R}^T(T_i, t_j^g) = \mathbb{I}^T(t_j^g \in T_i) \quad (6)$$

where $\mathbb{R}^T(T_i, t_j^g)$ returns 1 if the condition holds or 0 otherwise. Therefore, the *extended precision for top-P cascades* against news reports E is defined as follows.

$$EP@P = \frac{1}{P} \sum_{i=1}^P \sum_{j=1}^{|E|} \mathbb{R}^S(R_i, r_j^g) \times \mathbb{R}^T(t_j^g \in T_i) \quad (7)$$

A cascade is spatially and temporally accurate if it overlaps with more ground-truth without excessively expands its spatio-temporal scope. Higher EP indicates that higher-ranked cascades highly overlap with traffic news.

Comparison of Approaches. We perform congestion cascades identification using (1) CONN: connectivity-based, and (2)

TABLE II: Two Reported Traffic News on 2016-07-20

GT1	16:38	AYE	To MCE	After Exit A with congestion till Exit B
GT2	16:42	AYE	To MCE	After Exit A with congestion till Exit B

TABLE III: Extended Precision (EP@5) with $c=0.1$ and $\gamma=0.3$

P	CONN	DENS	Unified($\varphi=1.0, K=5, \epsilon=0.3$)
1	0.0	0.25	0.4
2	0.13	0.27	0.35
3	0.14	0.23	0.28
4	0.14	0.21	0.23
5	0.13	0.17	0.18

DENS: attribute-based, and (3) Unified: unified approaches. Based on five days of bus trajectory data and traffic news in July, we report the accuracy of congestion cascades identified by the examined approaches in Table III. Firstly, Unified has the best EP compared to other approaches. In particular, Unified yields the highest EP (0.4) at top-1 position. Secondly, as P increases, the EP of Unified decreases but remains the best. Thirdly, DENS outperforms CONN but it is still inferior to Unified, showing the strength of Unified which factors in both spatio-temporal connectivity and attribute coherence.

D. Case Study

To validate the practical value of BTCl, we further analyze cases of the identified congestion cascades on 2016-07-20 against the news reports summarized in Table II. Table IV shows the set of congestion cascades obtained with the parameter settings ($c=0.1, \gamma=0.3, \varphi=1.0, K=15, \epsilon=0.3$) in Unified approach. Each row provides detail of an identified congestion cascade, including number of c -segments ($cSeg$), number of connected components (nCC), number of impacted bus stations ($uStn$), number of unique time points ($uTpt$), temporal scope (T), average speed ($mSpd$), average membership ($mMem$), graph density (D), average congestion scores ($wScr$), the ranking metric ($D \times wScr$), and a flag for “News Hit”. In addition, we show some speed distribution of some cascades in Figure 6, where the curves plot the historical speed distribution from H and the blue histograms show the deviation of observed speed values. Among the top-5 cascades CG1 and CG2 overlap with the reported traffic news. Between them, CG1 overlaps with both GT1 and GT2 for around 2325 meters, both with 0.63 spatial coverage during [16:31-17:12]. Figure 6(a) and Figure 6(b) also show that both CG1 and CG2 have obvious deviation in observed bus speed toward the slow end against the historical norms. On the other hand, CG3, CG4 and CG5 are not reported in the news but the evidence of these congestions is very strong. For example, CG5⁶ consists of 593 c -segments with high congestion scores, which are spatially and temporally close to each other (involving 13 bus stations during [6:27-9:25]). Figure 6(c) also indicates serious deviation towards the very slow speed against the norm. The average historical speed within the same spatio-temporal scope of CG5 is 21.84 km/hr, whereas the average observed speed in CG5 drops to 12.97 km/hr with 77% of the speed values slower than 20 km/hr.

Spatio-Temporal Closeness. Beyond extended precision, we empirically evaluate the spatial and temporal closeness of c -

⁵The traffic news are reported at <https://twitter.com/ltattrafficnews>.

⁶Due to space constraint, we skip the discussion of CG3 and CG4.

TABLE IV: Identified Congestion Cascades by Unified ($c=0.1, \gamma=0.3, \varphi=1.0, K=15, \epsilon=0.3$) using Bus Trajectories on 2016-07-20

Rank	Cascade	cSeg	nCC	uStn	uTpt	T	mSpd	mMem	D	wScr	D × wScr	“News Hit”
1	CG1	58	1	8	42	[16:31-17:12]	15.27	0.485	0.83	0.543	0.451	Yes
2	CG2	13	1	8	29	[16:11-16:39]	15.31	0.956	0.86	0.492	0.423	Yes
3	CG3	27	1	4	48	[18:01-18:48]	10.59	1.000	0.61	0.559	0.341	No
4	CG4	250	1	10	88	[9:12-10:39]	12.05	0.349	0.32	0.580	0.186	No
5	CG5	593	1	13	179	[6:27-9:25]	12.98	0.465	0.14	0.586	0.082	No

TABLE V: Identified Congestion Cascades by DENS($c=0.1, \gamma=0.3, K=15$) using Bus Trajectories on 2016-07-20

Rank	Cascade	cSeg	nCC	uStn	uTpt	T	mSpd	mMem	D	wScr	D × wScr	“News Hit”
1	DG1	16	2	4	44	[10:19-10:34][10:36-11:03]	10.2	1.0	0.77	0.578	0.445	No
2	DG2	43	2	10	62	[16:11-17:12]	11.9	1.0	0.58	0.530	0.307	Yes
3	DG3	16	1	9	46	[10:19-11:04]	10.2	1.0	0.50	0.607	0.303	No
4	DG4	16	3	10	44	[10:19-11:02]	9.4	1.0	0.42	0.622	0.261	No
5	DG5	31	2	5	90	[8:54-9:15][9:23-10:30]	9.0	1.0	0.35	0.593	0.208	No

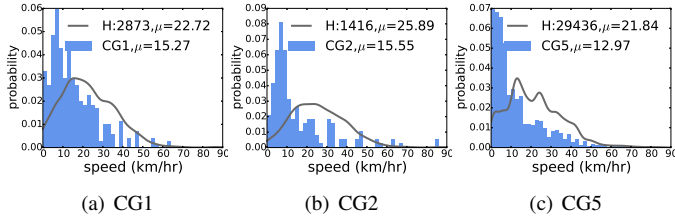


Fig. 6: Speed Distribution of CG1, CG2, and CG5 in Table IV

segments returned by each of the top 5 identified congestion cascades. We observe that each identified cascade in Table IV consists of c -segments that are east-bound and spatially nearby. For example, CG1, CG2 and CG5 center around the upper left stretch of the highway during different time periods. On the other hand, we observe that several cascades have c -segments temporally close to one another and the cascades occur during short periods of time as shown in Table IV. For example, CG1 has a short and continuous temporal scope [16:31-17:12]. While CG5 is a prolonged congestion, the congestion in the involved segments is particularly serious for the long period. **Comparison with DENS.** Table V shows the top 5 identified congestion cascades with the parameter settings ($c=0.1, \gamma=0.3, K=15$) in DENS approach. Compared to Unified approach, the spatio-temporal connectivity of c -segments in each of the top 5 identified congestion cascades returned by DENS is sacrificed. For example, DG1 and DG5 in Table V are discontinuous in time (T). Even though DG2 and DG4 have continuous temporal scopes, they are disconnected into 2 and 3 connected components (nCC) in space, respectively. The common disconnectivity in cascades returned by DENS suggests that DENS may not be a practical solution, as the result is not aligned with our intuition.

E. Sensitivity Study of Parameters

Impact of Lower Tail Threshold. Due to the lack of ground-truth in congested and non-congested segments, we have to empirically decide the congestion score threshold γ to extract c -segments. The higher γ is, the bar for being considered as a c -segment is higher, and thus less qualified c -segments. Table VI shows some structural properties of c -segment graphs extracted from the selected five days in July dataset, including average number of c -segments (\overline{cSeg}), average number of edges (\overline{E}), average graph density (\overline{D}), and average number of connected components (\overline{nCC}). As γ increases, the number

TABLE VI: C-Segment Network: $c=0.1$ with varying γ

γ	\overline{cSeg}	\overline{E}	\overline{D}	\overline{nCC}
0.1	697.4	39495.2	0.094	24.8
0.3	450.0	28351.2	0.143	12.8
0.5	236.6	14644.4	0.433	5.4
0.7	72.2	2643.2	0.242	2.4
0.9	2.4	6.8	0.062	0.6

TABLE VII: Unified($c=0.1, \gamma=0.3$) with varying K and ϵ

K	\overline{nCC}	\overline{uStn}	\overline{uTpt}	$\overline{D \times wScr}$	$\overline{EP@3}$
5	1.5	6.5	41.9	0.47	0.28
10	1.4	6.5	40.5	0.45	0.12
15	1.4	6.4	38.9	0.44	0.12
20	1.3	6.4	38.0	0.43	0.12
ϵ	\overline{nCC}	\overline{uStn}	\overline{uTpt}	$\overline{D \times wScr}$	$\overline{EP@3}$
0.0	1.5	6.5	41.9	0.58	0.28
0.3	1.5	6.5	41.9	0.58	0.28
0.6	1.4	6.5	39.0	0.57	0.2
0.9	1.3	4.8	26.9	0.6	0.0

of extracted c -segments and the size of these c -segments (in terms of \overline{cSeg} and \overline{E}) decrease. As a result, \overline{nCC} decreases. Note that γ controls the degree for segments to be considered as congested, but not necessarily the density of the resultant c -segment network. We set $\gamma=0.3$ for all experiments.

Impact of Number of Clusters. K controls the number of resultant cascades. A greater K results in more cascades. Table VII reports the average number of connected components (\overline{nCC}), average number of impacted bus stations (\overline{uStn}), average number of unique time points (\overline{uTpt}), and average ranking scores ($\overline{D \times wScr}$) of resultant cascades extracted from the selected five days in July dataset. Based on the five selected days of bus trajectory data and reported congestion news in July corresponding to various K , we observe that $K=5$ results in highest average $\overline{EP@3}$ in Table VII.

Impact of Membership Threshold. ϵ controls the strength of a c -segment associated to a cascade. With a higher ϵ only the c -segments strongly associated with a cascade is qualified as its member. Table VII shows how the resultant cascades behave with ϵ varied from 0.0 to 0.9. As ϵ increases, cascades are more concentrated spatially and temporally (smaller \overline{uStn} and \overline{uTpt}), stronger engagement (greater $\overline{D \times wScr}$). When ϵ is too strict (e.g., 0.9), the resultant cascades even exclude congestions reported in traffic news (i.e., $\overline{EP@3}=0.0$). In our work, we empirically set $\epsilon=0.3$ as it returns results consistent with the reported traffic congestion news with the best accuracy $\overline{EP@3}=0.28$.

VI. RELATED WORKS

Attributed Graph Clustering. There are many works on attributed graph clustering reported in literature, which can be categorized into two classes, distance-based [5][6] and model-based [7][3]. Distance-based approaches typically design a distance measure to fuse structural and attribute information and then apply standard clustering techniques (e.g., K-Medoids, spectral clustering) for attributed graph partitioning. Model-based approaches formulate a joint modelling of the edge connections and vertex attributes and use the model to infer the optimal clustering that best explains the attribute values and edge patterns. For example, Sun et al. propose a probabilistic generative model, a soft clustering solution that can handle various heterogeneous networks with categorical/numeric attributes and binary/weighted edges [3].

Traffic Anomaly Detection. The problem of detecting traffic anomalies has attracted considerable attention recently [8][9][10][11][12][13]. Some detect individual transport links with observed anomalous change of traffic flow [9], while some discovers anomalies with arbitrary spatio-temporal scope [11][13]. Zheng et al. in [13] propose to detect collective anomalies, where each anomaly refers to a collection of nearby regions that are anomalous during a few consecutive time intervals. They integrate multiple signals such as taxi flow and social media to determine anomalies with spatio-temporal scope. However, the structure properties and traffic health status of detected anomalies remain undiscovered. Moreover, those detected anomalies do not include congestion cascades. Some work relies on unique data sources (e.g., incident data) for impact modeling and prediction of incidents [11].

Traffic Condition Estimation. In traffic engineering, many research have been devoted to estimate traffic conditions on a road network [14][15][16][17]. These works rely on traffic sensors, loop detectors, cameras, and other instructions to obtain real-time traffic data to estimate vehicle speed, traffic density, and volume. Recently, research have also been carried out to utilizing Twitter as a new data source for detection [18][19][20] or visualization [21][22] of traffic events. Among multi-typed traffic events (e.g., congestion, accident, road construction and so on), only some works focus on studying traffic congestions estimation [19][20]. Due to inherent sparsity and low resolution of geographic locations in Twitter data, Wang et al. [20] propose a coupled matrix and tensor factorization model to effectively integrate rich information for traffic congestion estimation.

VII. CONCLUSIONS

We address a novel problem of identifying congestion cascades from sampled vehicle trajectories. To uncover congestion cascades, we propose *Bus Trajectory based Congestion Identification (BTCI)* framework that consists of two major components: (1) congested segment extraction, and (2) congestion cascade clustering. We first statistically capture the normal traffic health on segments from historical vehicle trajectory data and then we propose a statistics-based method to quantify a congestion score against the norm of traffic for segments during the time period of interest. Next, we propose to

aggregate congested segments into traffic congestion cascades by considering the spatio-temporal closeness and attribute coherence of segments in a cascade. Experimentation using 11.8 million bus transitions shows that the BTCI framework can effectively identify congestion cascades.

ACKNOWLEDGMENT

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative. Wang-Chien Lee's work was supported in part by National Science Foundation grant IIS-1717084.

REFERENCES

- [1] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, 1962.
- [2] A. Hinneburg, D. A. Keim *et al.*, "An efficient approach to clustering in large multimedia databases with noise," in *SIGKDD*, 1998.
- [3] Y. Sun, C. C. Aggarwal, and J. Han, "Relation strength-aware clustering of heterogeneous information networks with incomplete attributes," *VLDB Endowment*, vol. 5, no. 5, 2012.
- [4] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *SDM*, 2007.
- [5] Y. Zhou and L. Liu, "Social influence based clustering of heterogeneous information networks," in *SIGKDD*, 2013.
- [6] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," *VLDB Endowment*, vol. 2, no. 1, 2009.
- [7] Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng, "Gbagc: A general bayesian framework for attributed graph clustering," *TKDD*, vol. 9, no. 1, 2014.
- [8] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing, "Discovering spatio-temporal causal interactions in traffic data streams," in *SIGKDD*, 2011.
- [9] S. Chawla, Y. Zheng, and J. Hu, "Inferring the root cause in road traffic anomalies," in *ICDM*, 2012.
- [10] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *SIGSPATIAL*, 2013.
- [11] B. Pan, U. Demiryurek, C. Shahabi, and C. Gupta, "Forecasting spatiotemporal impact of traffic incidents on road networks," in *ICDM*, 2013.
- [12] S. Liu, Y. Liu, L. Ni, M. Li, and J. Fan, "Detecting crowdedness spot in city transportation," *Vehicular Technology*, vol. 62, 2013.
- [13] Y. Zheng, H. Zhang, and Y. Yu, "Detecting collective anomalies from multiple spatio-temporal datasets across different domains," in *SIGSPATIAL*, 2015.
- [14] L. Muñoz, X. Sun, R. Horowitz, and L. Alvarez, "Traffic density estimation with the cell transmission model," in *American Control Conference*, vol. 5, 2003.
- [15] F. Porikli and X. Li, "Traffic congestion estimation using hmm models without vehicle tracking," in *Intelligent Vehicles Symposium*, 2004.
- [16] W. Pattara-Atikom, P. Pongpaibool, and S. Thajchayapong, "Estimating road traffic congestion using vehicle velocity," in *ITS Telecommunications*, 2006.
- [17] C. De Fabritiis, R. Ragona, and G. Valenti, "Traffic estimation and prediction based on real time floating car data," in *ITSC*, 2008.
- [18] E. M. Daly, F. Lecue, and V. Bicer, "Westland row why so slow?: fusing social media and linked data sources for understanding real-time traffic conditions," in *IUI*, 2013.
- [19] P.-T. Chen, F. Chen, and Z. Qian, "Road traffic congestion monitoring in social media with hinge-loss markov random fields," in *ICDM*, 2014.
- [20] S. Wang, L. He, L. Stenneth, P. S. Yu, and Z. Li, "Citywide traffic congestion estimation with social media," in *SIGSPATIAL*, 2015.
- [21] M. Liu, K. Fu, C.-T. Lu, G. Chen, and H. Wang, "A search and summary application for traffic events detection based on twitter data," in *SIGSPATIAL*, 2014.
- [22] E. D'Andrea, P. Ducange, B. Lazzarini, and F. Marcelloni, "Real-time detection of traffic from twitter stream analysis," *Intelligent Transportation Systems*, vol. 16, no. 4, 2015.