Frequent Pattern Analysis of the Roadside Safety Devices Related On-road Crashes

Dr. Yunpeng (Jack) Zhang

Assistant Professor

Department of Information & Logistics Technology

University of Houston

yzhan226@central.uh.edu

University of Houston

- A public research university in Houston, Texas.
- Founded in 1927, University of Houston(UH) is the flagship institution of the University of Houston System.
- The third-largest university in Texas with over 46,000 students.
- R1: Doctoral Universities – Very high research activity.
- Offers more than 282 degree programs through its academic colleges.
- Conducts $150 million annually in research and operates more than 40 research centers and institutes on campus.
- Awarding more than 9,000 degrees annually, UH's alumni base exceeds 260,000.

Dr. Zhang’s Research overview

Cybersecurity

Software Engineering

Intrusion Detection
Chaos Cryptography
DNA Cryptography
Access Control

Human Computer Interface
Code Testing

Data Science

Machine Learning
Data Mining

Dr. Zhang’s Website
https://uh.edu/technology/departments/ilt/people/faculty/?l=zhang&f=yunpeng

Dr. Zhang’s Lab, Network Software and System Lab
https://network-software-and-system-lab.github.io/index

Topics Covered

- Motivation
- Summary
- Literature Review
- Crash Data Used
- Measures used
- Frequent pattern algorithms
- Results
- Conclusion
- Future Work and Acknowledgements

Motivation

- FARS of U.S. NHTSA in 2018 reported 36,560 nationwide highway fatalities with fatality rate of 1.13 per 100 million vehicles miles.
- Causes:
  - 1st most : Human factors: drivers’ actions (e.g. speeding) or conditions (e.g., alcohol or drug effects)
  - 2nd most : Roadway factors: roadway design, use of traffic control devices, and land-use configurations
  - others : vehicle and traffic factors, environmental factors
- Roadside safety control devices (related to Roadway factors):
  - installed on roadsides to reduce the risk of serious and fatal injuries to motorist’s inadvertent road departures
  - performance criteria are detailed in the Manual for Assessing Safety Hardware (MASH) standards but are based on full-scale crash testing evaluation under ideal site conditions with carefully controlled conditions
  - in-service performance evaluation (ISPE) as the final step in truly evaluating roadside hardware

Motivation

- List of safety devices in MASH 2016 includes longitudinal barriers, terminals, crash cushions, support structure, work zone attenuation, and channelizers, drainage features, geometric features, and other devices.
- Fig. 1 demonstrates six types of roadside safety devices :(a) concrete traffic barrier, (b) low-tension cable median barriers, (c) W-beam guardrail, (d) concrete railing, (e) metal and concrete railing, (f) transition of bridge rail end.

(a) (b) (c) (d) (e) (f)

Figure 1. Typical roadside safety devices on highway and roadway

Motivation

- Differences between field performance and crash test results caused by
  - field impact
  - maintenance conditions
- To do the ISPE, In this research, the **frequent pattern-based data mining approach** is adopted to characterize the associations between roadside safety devices and different types of crashes
- Results of this study will prioritize the performance of traffic control devices based on their associated crashes, so as to improve the design, testing, and maintenance of roadway traffic control devices for the benefits of safety enhancement

Summary

- Apriori and FP-Growth frequent pattern mining algorithms were applied to characterize the relationship between roadside safety devices and on-road crashes
- We discuss the flow chart and pseudo-codes of the Apriori and FP-Growth algorithms and the calculation equations of the evaluation parameters
- **Ten-year roadway crash data** from TxDOT database was collected, which contains various crash influencing factors and six roadside safety devices
- Raw data was fitted into a set of lists as part of the input, along with the minsup for the frequent pattern algorithm
- Proper algorithm was selected based on the optimal running time.
- Through data analysis, the trends of the ten-year crash data were found.
- Associations between the crashes and their influencing factors were elaborated through the mining of frequent patterns.

Literature Review

- Frequent pattern mining was originally introduced for mining of association rules for market-basket analysis
- Popular Frequent pattern algorithms
  - Apriori algorithm(1994)
  - Frequent Pattern (FP)-growth algorithm(2004)
- Measures used:
  - Support and Confidence
  - Correlation measure: valuates the correlation between two itemsets by comparing their separate and union occurrences.
  - Other measure available: all confidence, max confidence, and cosine measures, etc.

Literature Review

- Similar Research
  - Juan et al.(2008) : FP-growth algorithm to process traffic violation data in ITS
  - Glatz et al. (2014):frequent pattern mining method to visualize traffic network data with communication logs
  - Das and Sun(2014): association rule method to discover hidden patterns in rainy weather crash data
  - Kumar et al.(2017): K-Modes clustering approach to categorize and analyze accident data for heterogeneity reduction
  - Lin et al.(2017): FP-growth based variable selection method for real-time risk prediction models for traffic accidents
  - Xia et al. (2018): MapReduce-based Parallel Frequent Pattern growth (MR-PFP) to analyze characteristics in taxi operation

Crash Data Used

- Crash data
  - collected from TxDOT
  - ten-years (January 1st, 2010 to December 31st, 2019)
  - 5,629,779 crashes
  - 172 features (information of crash, unit, person, charges, primary person, endorsements, restrictions, and damages, etc.)
  - Some Feature codes are shown below
- transformed into a set of lists so that, each crash record including its factors, is an inner list within the outer list of all records

Crash Data Used

Table 1. Typical Variables with Codes in Texas Crash Database

Safety Device	**Weather Condition**	**Light Condition**	**Surface Condition**	**Day of Week**	**Crash Speed Limit (mph)**
23- guardrail	0- unknown	0- unknown	0- unknown	Monday	-1(No data)
28- work zone barricade, cones, signs or material	2- rain	1- daylight	1- dry	Tuesday	5
39- median barrier (concrete or cable)	3- sleet/hail	2- dawn	2- wet	Wednesday	10
40- end of bridge (abutment or rail end)	4- snow	3- dark, not lighted	3- standing water	Thursday	15
41- side of bridge (bridge rail)	5- fog	4- dark, lighted	5- slush	Friday	20
56- concrete traffic barrier (not in median)	6- blowing sand/snow	5- dusk	6- ice	Saturday	25
	7- severe crosswinds	6- dark, unknown lighting	8- other	Sunday	30
	8- other	8- other	9- snow		35
	11- clear		10- sand, mud, dirt		…
	12- cloudy				80

Measures used

- Support s(C, D) : provides the scale of the crash occurring on an influencing item, which is calculated from the number of crashes under the influencing item divide by the total crash number
- Confidence c(C, D) :likelihood of an item occurs if another item happened, which is calculated from the support of two events happen together divide by the support of the single event
- LIFT l(C, D): illustrates the increase in a crash when another item happened, which is calculated from the support of two events that happen together divide by the grade of the supports of the two single events
- The Support, Confidence, and the interestingness measurement LIFT can be calculated using

$$s(C,D) = s(C \cup D) = \frac{n(C \cup D)}{n(T)}$$

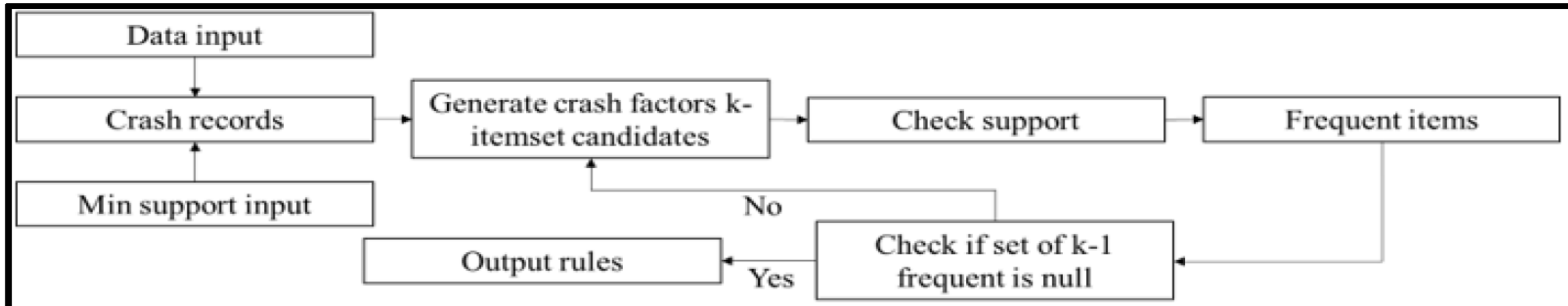
$$c(C,D) = \frac{s(C \cup D)}{s(C)}$$

$$l(C,D) = \frac{c(C \cup D)}{s(D)} = \frac{s(C \cup D)}{s(C) * s(D)}$$

- where,
  - s(C, D): the Support for crash C and device D occurring together, ranging (0, 1);
  - n(C, D): the number of events when C and D occurring together;
  - n(T): the number of total events;
  - c(C, D): the Confidence for event D to occur when event C occurs, ranging (0, 1);
  - l(C, D): the interestingness measurement LIFT (ranging (0, ∞)) for event D to occur when event C occurs, which tells how C and D are correlated;
    - if l(C, D) = 1, events C and D are independent;
    - if l(C, D) in (1, ∞), events C and D are positively correlated;
    - if l(C, D) in (0, 1), events C and D are negatively correlated.

Frequent pattern algorithms

- **Apriori algorithm** scans all possible itemsets and conducts all calculations
  - If the *Support* of the candidate itemset is greater than the *minsup*, the frequent items are recorded, and the process goes through the null test.

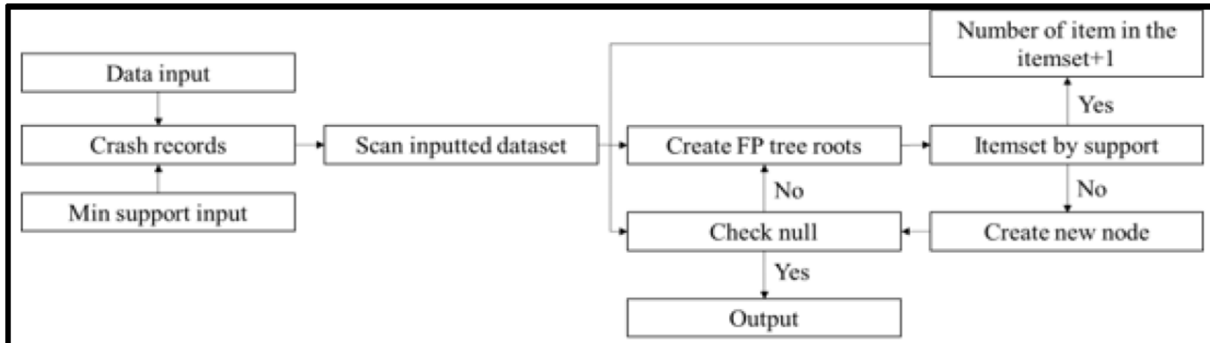



```
Cₖ: Candidate itemset of size k;
Fₖ: Frequent itemset of size k;
k := 1;/
Fₖ := {frequent items}; // frequent 1 − itemset
While (Fₖ! = ∅)// when Fₖ is non empty
      do { Cₖ₊₁ := candidates generated from Fₖ; // candidate generation
           Derive Fₖ₊₁ by counting candidates in Cₖ₊₁ with respect to TDB at min_support;
           k := k + 1}
return ∪ₖ Fₖ // return Fₖ generated at each level
```

Figure 3. The flow chart and pseudo-code of the Apriori algorithm [8]

- **FP-Growth algorithm** does not consider all possible itemsets. Includes below steps
  - creating the FP-Tree
  - applying the FP-Growth algorithm
- Which to use?
  - mining results of the Apriori and FP-Growth algorithms are the same
  - FP-growth algorithm runs faster when the settled *minsup* is under a specific range
  - If the *minsup* is relatively small, it would be more efficient to use the Apriori algorithm.
- To include impact of crash severity, Equivalent Property Damage Only (EPDO) weights of crashes are added as below

• Scale ID *K*:	Fatal injury (death within 30 days),	weight 568
• Scale ID *A*:	Suspected serious injury,	weight 30
• Scale ID *B*:	Suspected minor injury,	weight 11
• Scale ID *C*:	Possible injury,	weight 6
• Scale ID *O*:	No apparent injury,	weight 1




```
// FP-Tree
T: List of transactions;
min_sup: minimum support value;
define F[] = Ø;
for each transaction T_i in T
    for each item a_j in T_i
        do{ F[a_i] ++}
sort F[];
define null root r;
for each transaction T_i in T
   do{ T_i ordered by F
       call insert_tree(T_i, r)};
for each transaction T_i in T
end for
// insert_tree(T_i, r)
if r contains note N & N.item_name = T.item_name ;
  then N.count ++;
  else creat new node N;
       N.item − name = p.item − name;
       N.count ++;
       T.parent = r;
       point N.node_link to the node with same name
 end if
 if T! = Ø;
   then N.count ++;
   call insert_tree(T_i, r);
 end if
```

```
// FP-Growth(FP-Tree, Ø )
If tree contains single path P then
for nodes combination of path P noted as β
    do{ create pattern β ∪ α,
        support = min_sup of nodes in β}
else for each α_i in tree
        do{ create pattern β = α ∪ α_i
            support = α_i.support
            construct β conditional database
            construct β conditional Tree_β
            if Tree_β! = Ø
              then call FP − Growth(Tree_β, β)
            end if }
     end for
end if
```

Figurer 4. The flow chart and pseudo-code of the FP-Growth algorithm [25]

Results

- Among all safety devices,
  - nearly half (46.0%) crashes were associated with “Median Barriers”,
  - 28.3% and 19.5% crashes were related to “Guardrail” and “Concrete Traffic”.
  - Other safety devices were related to the rest of the 6.2% crashes.

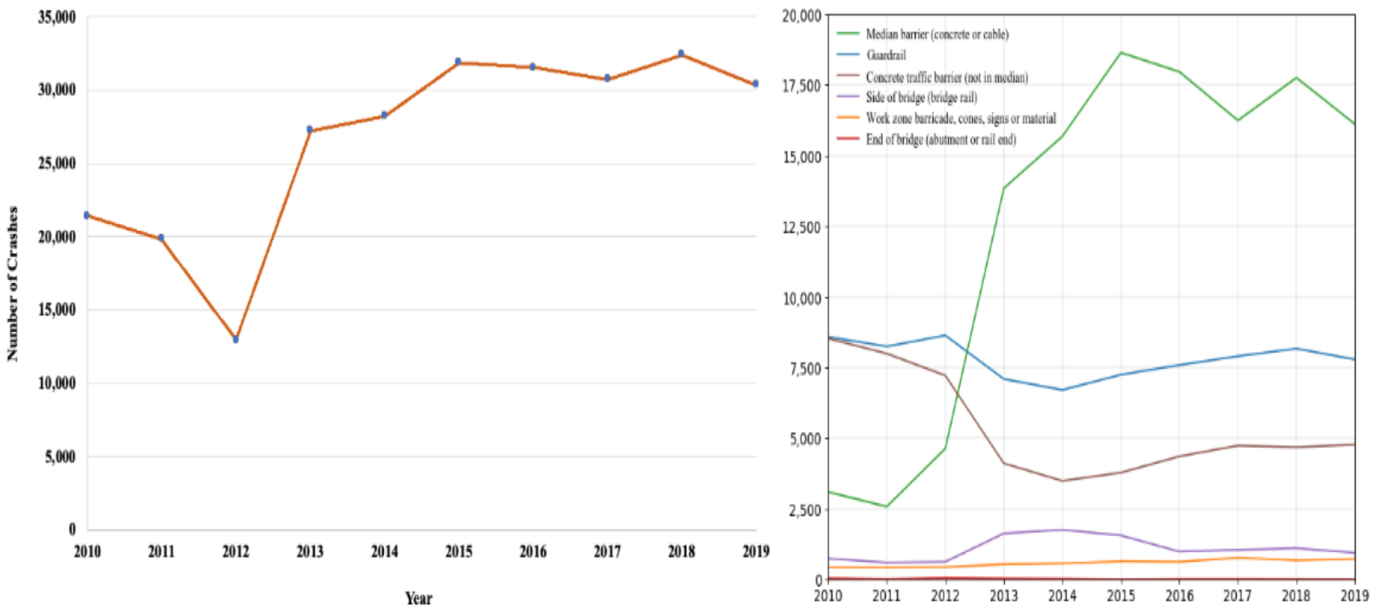


Figure 5. Total number of crashes associated with safety devices in Texas from 2010 to 2019

Table 2. Number of Ten Years Texas Crash Based on Types of Roadside Safety Devices

Types of Safety Device	Guardrail	Work Zone Barricade, Cones, Signs or Material	Median Barrier	End of Bridge	Side of Bridge	Concrete Traffic	Total Crash
2010	8,586	430	3,098	45	740	8,530	21,429
2011	8,247	426	2,578	20	600	7,996	19,867
2012	8,639	439	4,635	55	625	7,219	**12,973**
2013	7,095	536	13,845	38	1,631	4,110	27,255
2014	6,709	568	15,684	30	1,760	3,490	28,241
2015	7,247	643	18,646	11	1,561	3,781	31,889
2016	7,590	625	17,968	23	986	4,358	31,550
2017	7,903	769	16,246	23	1,049	4,738	30,728
2018	8,171	677	17,759	20	1,106	4,681	32,414
2019	7,786	727	16,114	13	943	4,775	30,358
Total	**77,973**	**5,840**	**126,573**	**278**	**11,001**	**53,678**	**266,704**
	28.3%	**2.1%**	**46.0%**	**0.1%**	**4.0%**	**19.5%**	**100.0%**

- To select the suitable algorithm (Apriori or FP-growth), average running times is considered as shown in Table 3 and Apriori algorithm was employed when the minsup > 10%. Otherwise, the FP-Growth algorithm was used.
- The Supports for safety devices as itemsets were as below and this result is consistent with the crash trend analysis
  - 46.0% for median barrier
  - 28.3% for guardrail
  - 19.5% for concrete traffic barrier
  - 4.0% for side of bridge
  - 2.1% for work zone barricade, cones, signs or material
  - 0.1% for the end of bridge
- When considering the safety devices as items , other factors such as the “Surface condition”, “Day of weeks”, “Crash speed limit”, “Weather condition”, and “Light condition” are used to get the confidence whose Support-Confidence plots are shown in Fig. 6.

Table 3. Running Time of Frequent Pattern Algorithms

minup	Running time per loop (10 runs, 100 loops each, mean ± std. dev.) Apriori	FP-Growth
60%	127 ms ± 4.61 ms	3.33 s ± 0.155 s
50%	146 ms ± 4.09 ms	4.11 s ± 0.285 s
40%	146 ms ± 2.89 ms	4.48 s ± 0.407 s
30%	156 ms ± 3.67 ms	4.52 s ± 0.709 s
20%	327 ms ± 6.98 ms	6.07 s ± 1.53 s
10%	3.72 s ± 0.111 s	5.84 s ± 0.550 s
5%	14.4 s ± 1.44 s	5.92 s ± 0.312 s
1%	154 s ± 33 s	5.28 s ± 0.614 s

Table 4. Frequent Itemsets Including Safety Devices

Itemsets Safety Devices	Weather condition	Light condition	Surface Condition	Day of Week	Crash speed limit (mph)	Support
Guardrail	Clear	Dark lighted	Dry	Sunday	60	0.15%
Median barrier	Clear	Daylight	Dry	Friday	65	0.25%
Side of bridge	Clear	Daylight	Dry	Wednesday	60	0.01%
	Clear	Daylight	Dry	Tuesday	70	0.01%
Concrete traffic barrier	Clear	Dark lighted	Dry	Sunday	60	0.18%

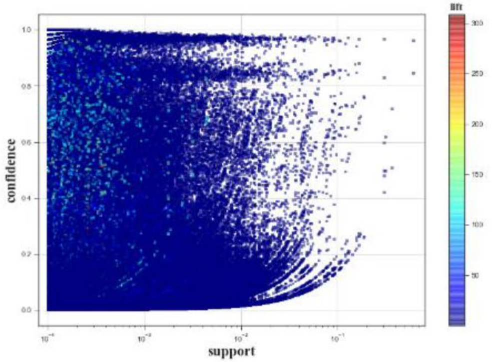


Figure 6. Confidence-Support relationship with the colors representing relevant LIFT values

- Support-confidence relation of each safety device is shown in Fig7
- Illustration of frequent itemsets with higher *Support* for each Safety Device is shown in Fig8
- While the Prioritized Safety Device ID Under Different Crash Severity is shown in Table 5

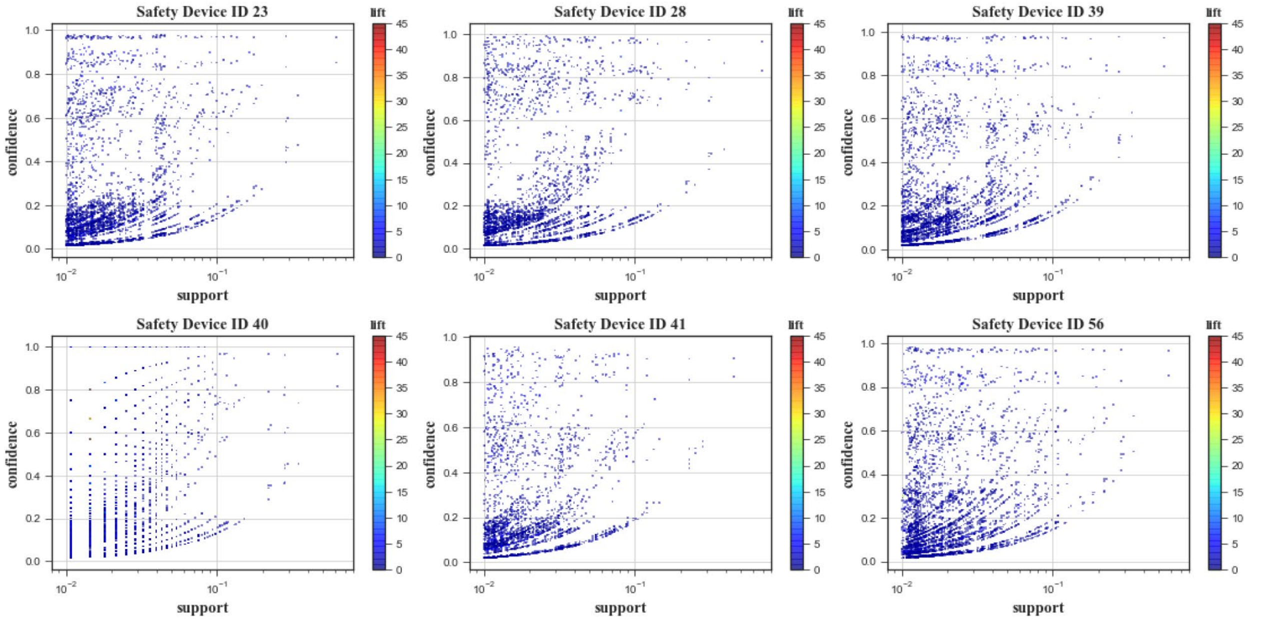


Figure 7. Support-confidence relation of each safety device

Table 5. The Prioritized Safety Device ID Under Different Crash Severity

Year	Killed/Fatal Injury (K)	Incapacitating Injury/Suspected Serious Injury (A)	Non-Incapacitating Injury (B)	Possible Injury (C)	Unknown/Not Injured (O)
Overall	23, 39, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	23, 39, 56, 28, 41, 40
2019	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 28, 41, 40
2018	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 28, 41, 40
2017	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 28, 41, 40
2016	23, 39, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 28, 41, 40
2015	23, 39, 41, 56, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 28, 41, 40
2014	39, 23, 41, 56, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40
2013	39, 23, 41, 56, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40	39, 23, 56, 41, 28, 40
2012	23, 39, 56, 41, 28, 40	23, 56, 39, 41, 28, 40	23, 56, 39, 41, 28, 40	56, 23, 39, 41, 28, 40	23, 56, 39, 41, 28, 40
2011	23, 56, 39, 41, 28, 40	23, 56, 39, 41, 28, 40	56, 23, 39, 41, 28, 40	56, 23, 39, 41, 28, 40	23, 56, 39, 28, 41, 40
2010	23, 56, 39, 41, 40, 28	23, 56, 39, 41, 28, 40	56, 23, 39, 41, 28, 40	56, 23, 39, 41, 28, 40	23, 56, 39, 41, 28, 40

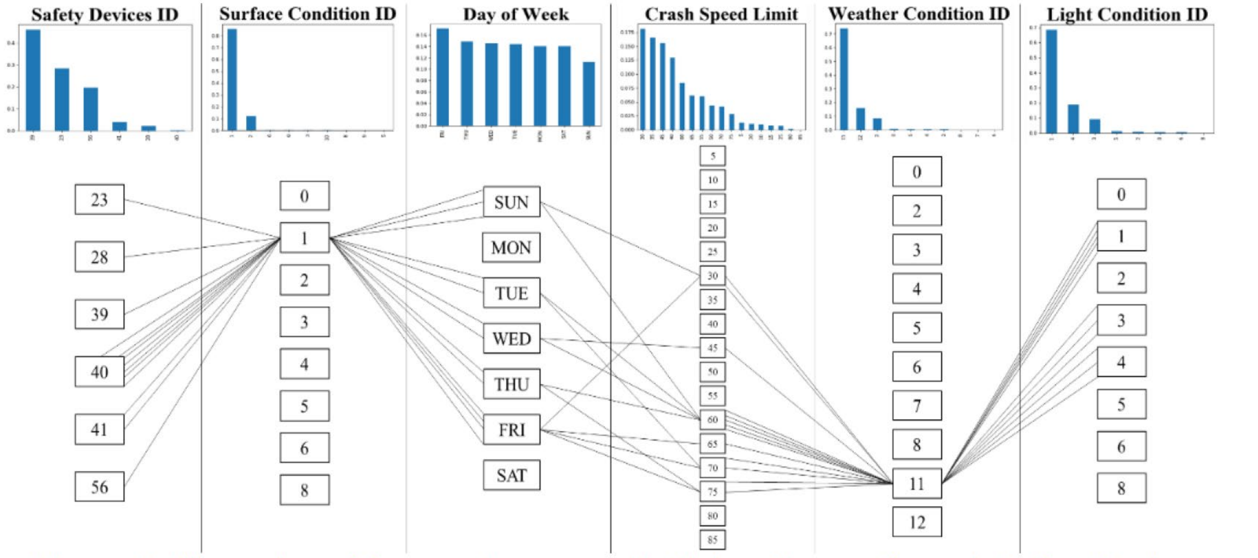


Figure 8. Illustration of frequent itemsets with higher *Support* for each Safety Device

- Table 6 and Table 7 show the **Prioritized Crash Severity Under Different Safety Device and Safety Device / Crash Severity EPDO Index when weights based on EPDO were added**

Table 6. The Prioritized Crash Severity Under Different Safety Device

Year	Guardrail (23)	Work Zone Barricade, Cones, Signs or Material (28)	Median Barrier (39)	End of Bridge (40)	Side of Bridge (41)	Concrete Traffic Barrier (56)
Overall	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **B, C, K, A**	O, C, B, A, K	O, C, B, A, K
2019	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **K, C, B, A**	O, C, B, A, K	O, C, B, A, K
2018	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **C, B, K, A**	O, C, B, A, K	O, C, B, A, K
2017	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **C, B, K, A**	O, C, B, A, K	O, C, B, A, K
2016	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **K, B, A, C**	O, C, B, A, K	O, C, B, A, K
2015	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **C, K, B, A**	O, C, B, A, K	O, C, B, A, K
2014	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **B, K, C, A**	O, C, B, A, K	O, C, B, A, K
2013	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **B, C,** A, K	O, C, B, A, K	O, C, B, A, K
2012	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **K, C, B, A**	O, **B, C,** A, K	O, C, B, A, K
2011	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **B, A, K, C**	O, C, B, **K, A**	O, C, B, A, K
2010	O, C, B, A, K	O, C, B, A, K	O, C, B, A, K	O, **B, K, C, A**	O, C, B, **K, A**	O, C, B, A, K

Table 7. The Safety Device / Crash Severity EPDO Index

Year	Guardrail	Work Zone Barricade, Cones, Signs or Material	Median Barrier	End of Bridge	Side of Bridge	Concrete Traffic Barrier
Overall	11.06	10.74	7.97	**68.31**	18.18	7.72
2019	9.11	8.74	7.74	**136.77**	15.81	6.97
2018	11.25	7.03	7.74	33.15	14.10	8.17
2017	10.81	8.09	7.65	29.04	15.44	8.43
2016	**12.52**	**16.91**	7.41	129.17	18.65	**8.70**
2015	12.28	7.01	7.00	54.82	17.17	7.99
2014	11.59	12.40	7.55	43.70	13.05	7.97
2013	10.68	13.06	8.97	51.66	19.29	7.91
2012	10.13	15.87	**11.76**	85.20	25.47	6.45
2011	10.67	9.90	11.61	5.35	26.93	7.65
2010	11.76	11.79	10.14	94.20	**29.14**	7.80

Conclusion

- Frequent pattern mining results suggest that crashes are likely to happen on dry surface pavement, in clear weather, and under daylight or dark light conditions
- No-injury crashes rank number one for all roadside devices, while the fatal crashes rank the last for roadside safety devices except for the “end of bridge” and the “side of bridge”.
- It is suggested that certain countermeasures and treatments shall be designed and implemented for the roadside device “end of bridge”, while the “side of bridge” shall be put on a “watch list”.
- The crash severities did not vary much within the 10 years.
  - The mildest crashes were related to the “concrete traffic barrier”
  - The harshest crashes were related to the “end of bridge”.
  - The average crash severities related to the roadside safety devices were likely to be severer than a suspected minor injury
- The safety device “media barrier”, while is related to 46.0% of the total crashes in Texas, the EPDO index of which is however generally very low.

Future Work

As a plan of the future work of this study, the design (color, reflection, etc.), length, year of service, and maintenance records of roadside devices will be included in the next phase studies.

Acknowledgements

- This research is supported in part by the Texas Department of Transportation (TxDOT) with project number 0-7018
- We appreciate the comments and support from the project manager Wade Odell and other members of the Project Management Committee

Q & A

Thank You!!