BLUE MARTINI SOFTWARE

Real World Performance Of Association Rule Algorith

Zijian Zheng, Ron Kohavi, and Llew Mason

{zijian,ronnyk,lmason}@bluemartini.com

Blue Martini Software

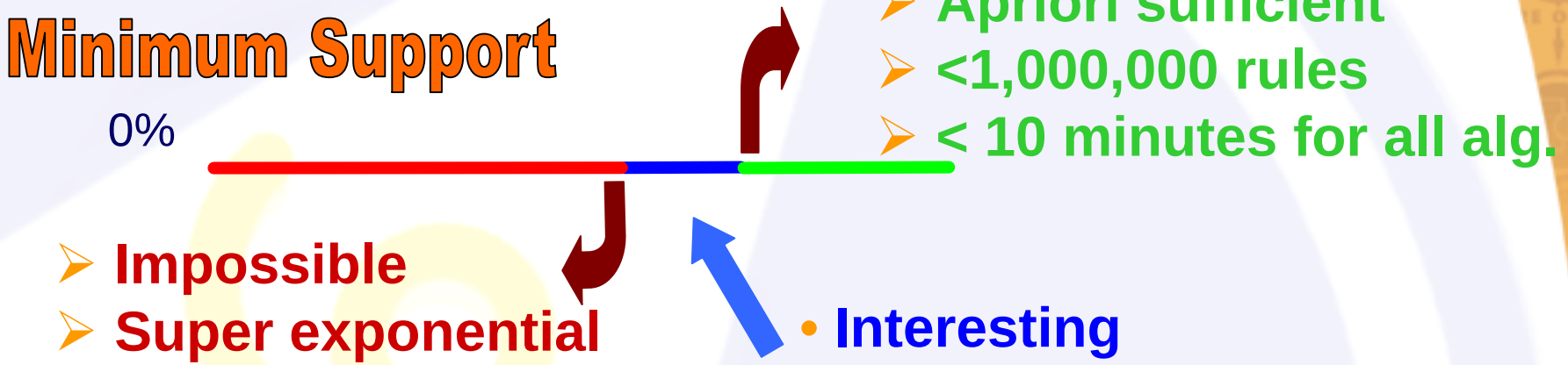
San Mateo, California

August 11, 2001

Performance Improvement Irrelevant

BLUE MARTINI SOFTWARE

- Very narrow min-sup range of interest
- Super exponential growth in number of rules on real world data



- Fast, but incorrect results

Artificial Improvement

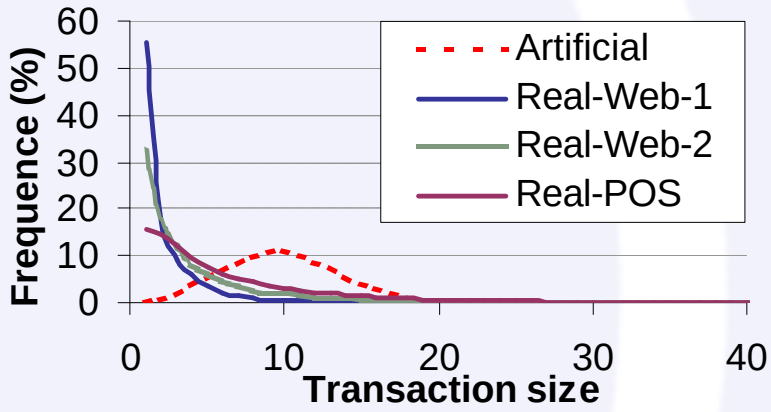
BLUE MARTINI SOFTWARE

- **Performance improvements on artificial data did not generalize to real world data**

Improvement over Apriori:

	Closet	Charm	FP-growth	Avg
Artificial	**1.9 x**	**2.8 x**	**12.1 x**	**3.1 x**
Real-POS	**1.2 x**	**1.1 x**	**0.8 x**	**1.0 x**

- **Are algorithms overfitting the artificial data?**



Background & Motivation

BLUE MARTINI SOFTWARE

- Association rule discovery:
  - Active research area
  - Good application potential
- Many promising new algorithms
- Each new algorithm has significant performance improvements
  - mainly based on results on IBM Almaden artificial datasets
- How do these algorithms perform on real-world data?

Datasets

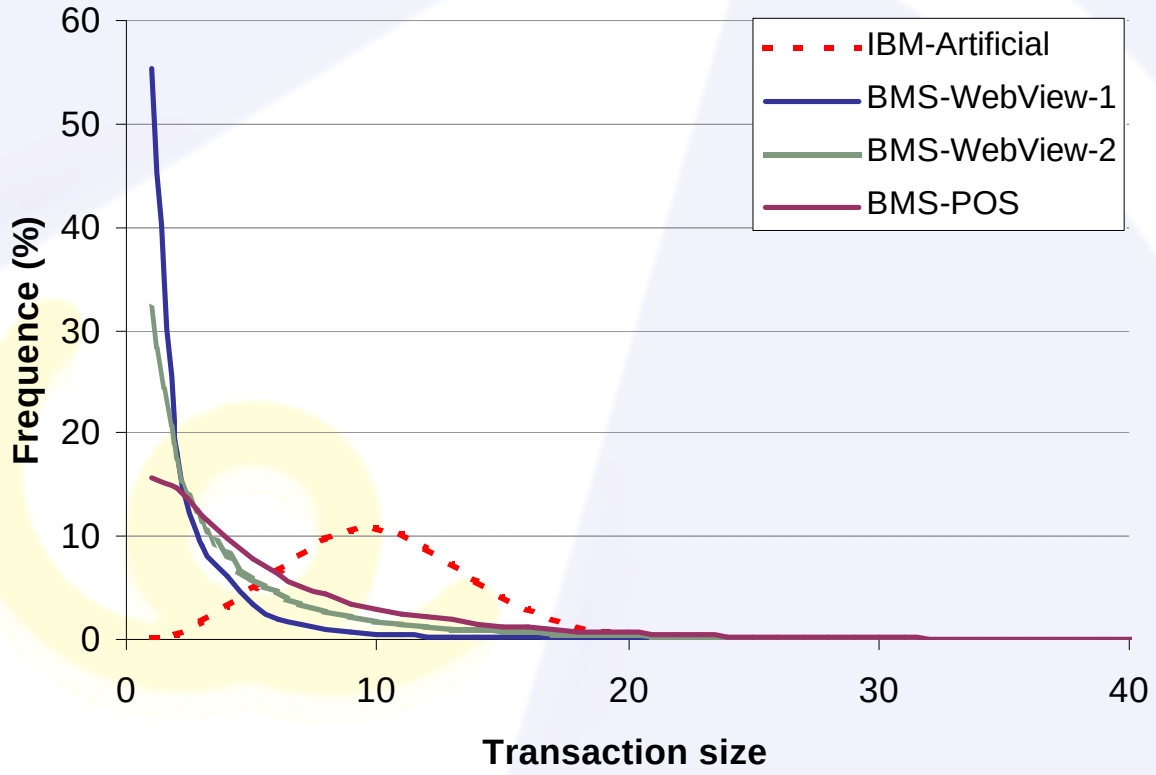
BLUE MARTINI SOFTWARE

	Transactions	Distinct Items	Maximum Transaction Size	Average Transaction Size
IBM-Artificial	100,000	870	29	10.1
BMS-POS	515,597	1,657	164	6.5
BMS-WebView-1	59,602	497	267	2.5
BMS-WebView-2	77,512	3,340	161	5.0

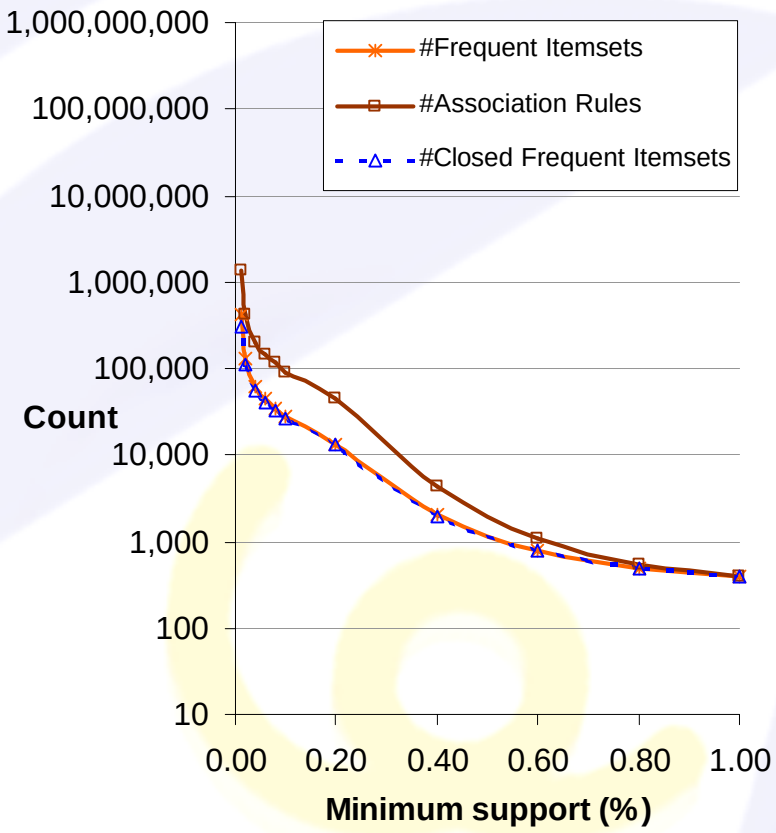
Datasets (Cont'd)

BLUE MARTINI SOFTWARE

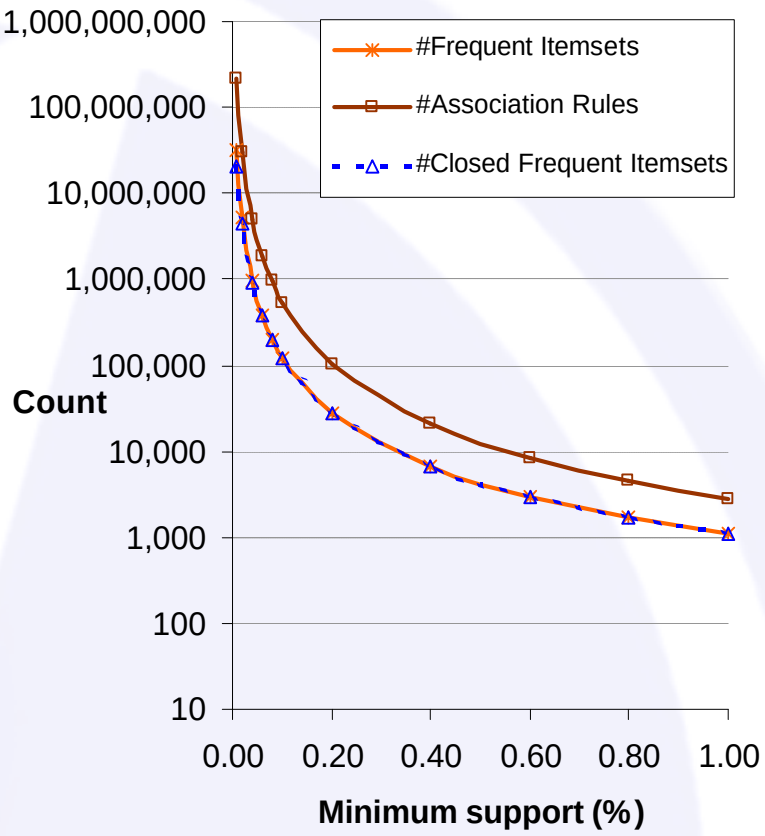
IBM artificial dataset is very different from the real-world datasets



Datasets (Cont'd)

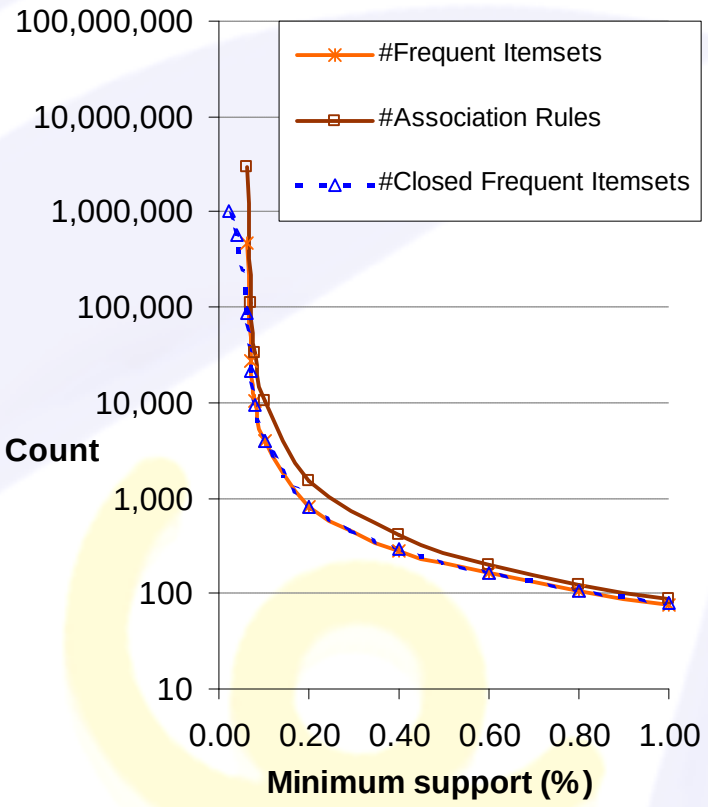


IBM-Artificial

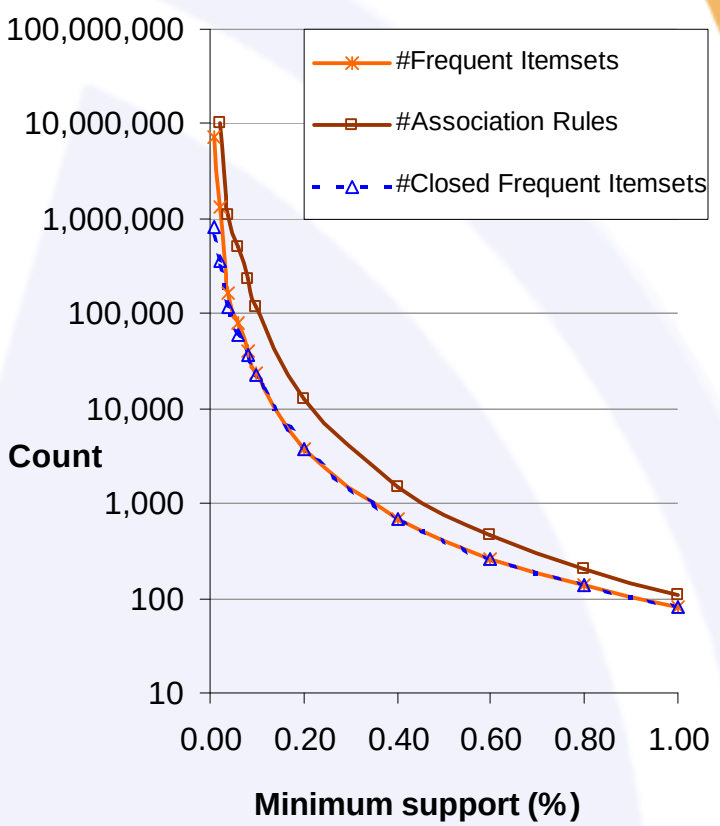


BMS-POS

Datasets (Cont'd)

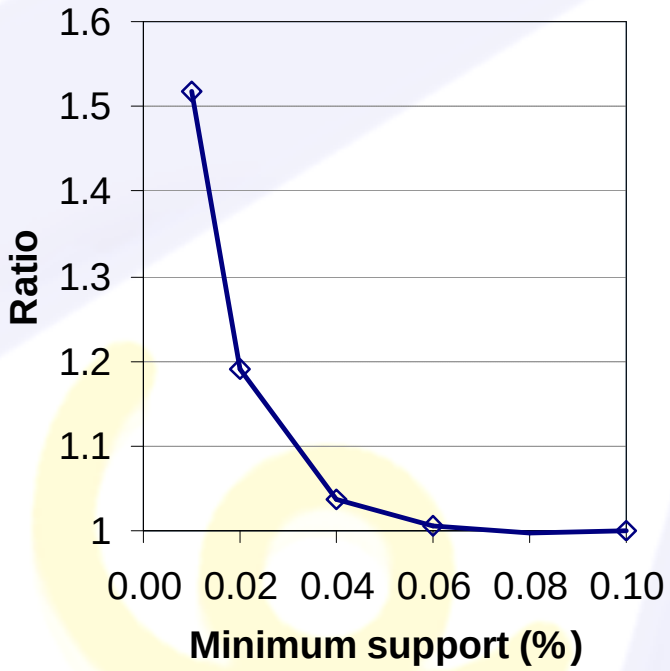


BMS-WebView-1

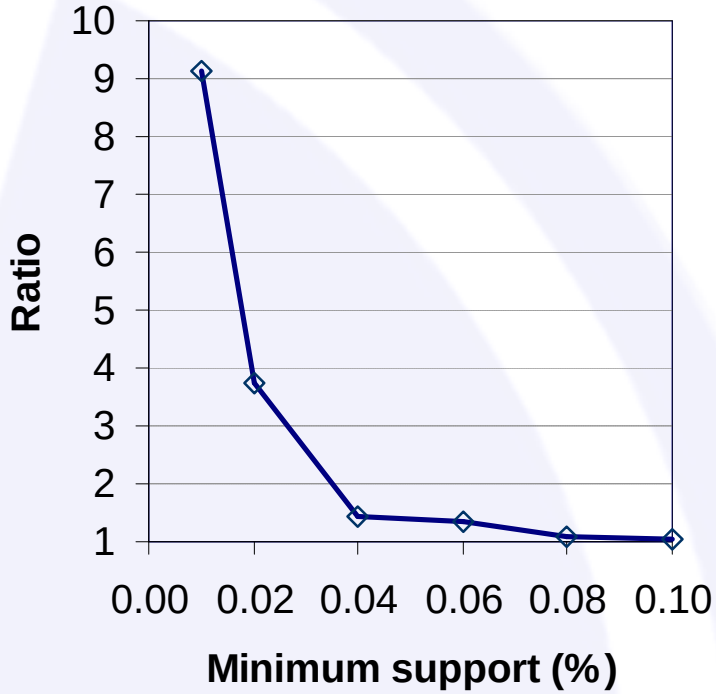


BMS-WebView-2

Ratio of #Freq itemsets over #Closed freq itemsets



BMS-POS



BMS-WebView-2

Experimental Setup

- Dual 550MHz Pentium III with 1 GB of memory
- Windows NT 4.0
- Measures: time (seconds) for generating frequent itemsets/association rules
- Minimum support: 1.00%, 0.80%, 0.60%, 0.40%, 0.20%, 0.10%, 0.08%, 0.06%, 0.04%, 0.02%, and 0.01%
- Minimum confidence: 0%

Comparison Algorith

BLUE MARTINI SOFTWARE

- **Apriori:** C. Borgelt's implementation
- **Charm:** M. Zaki
- **FP-growth:** J. Han's research group
- **Closet:** J. Han's research group
- **MagnumOpus :** G. Webb

Experimental Results

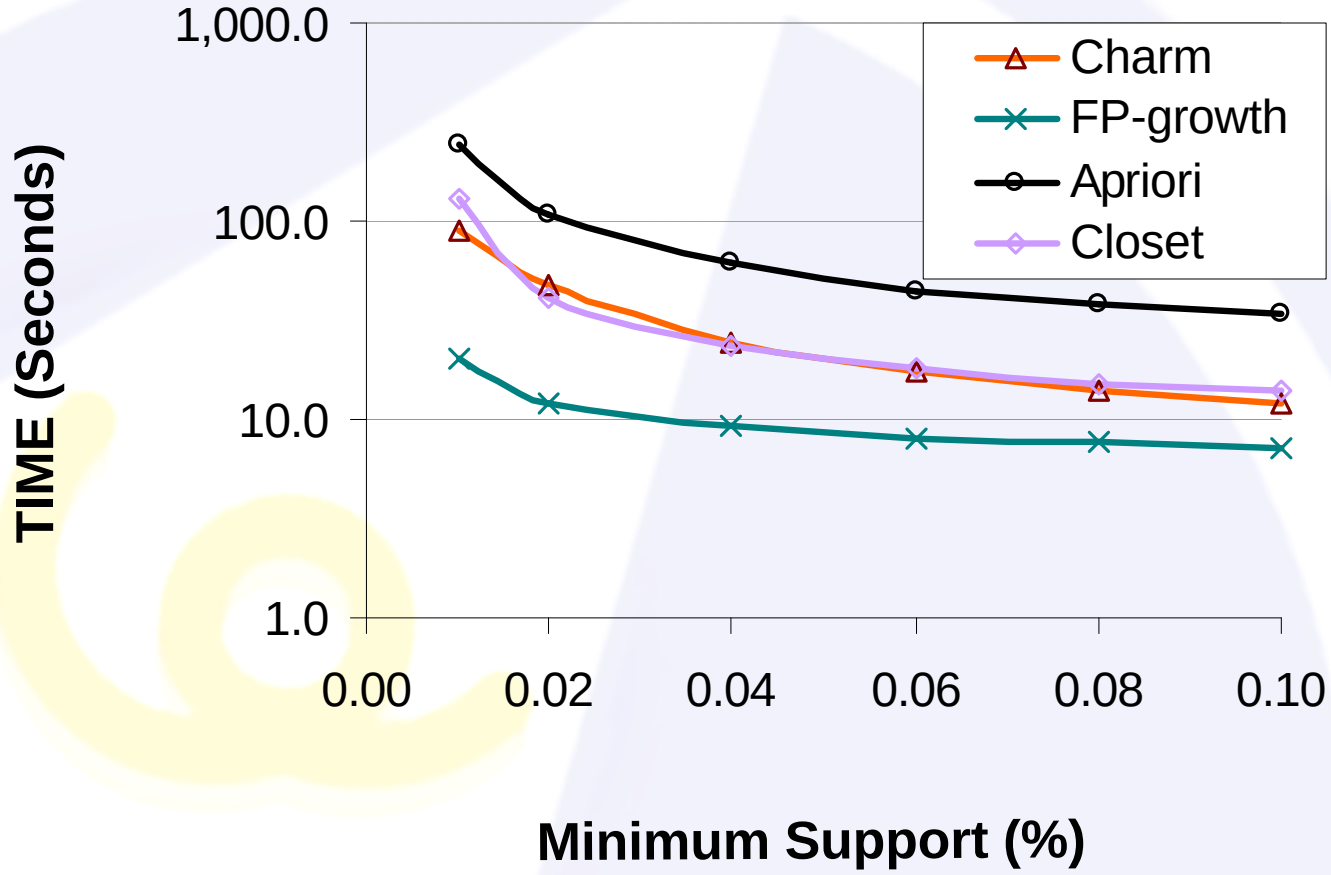
BLUE MARTINI SOFTWARE

Rankings (left is better) of the algorithms for generating frequent itemsets on the four datasets with high minimum supports and low minimum supports (Ap: Apriori, FP: FP-growth, Ch: Charm, Cl: Closet):

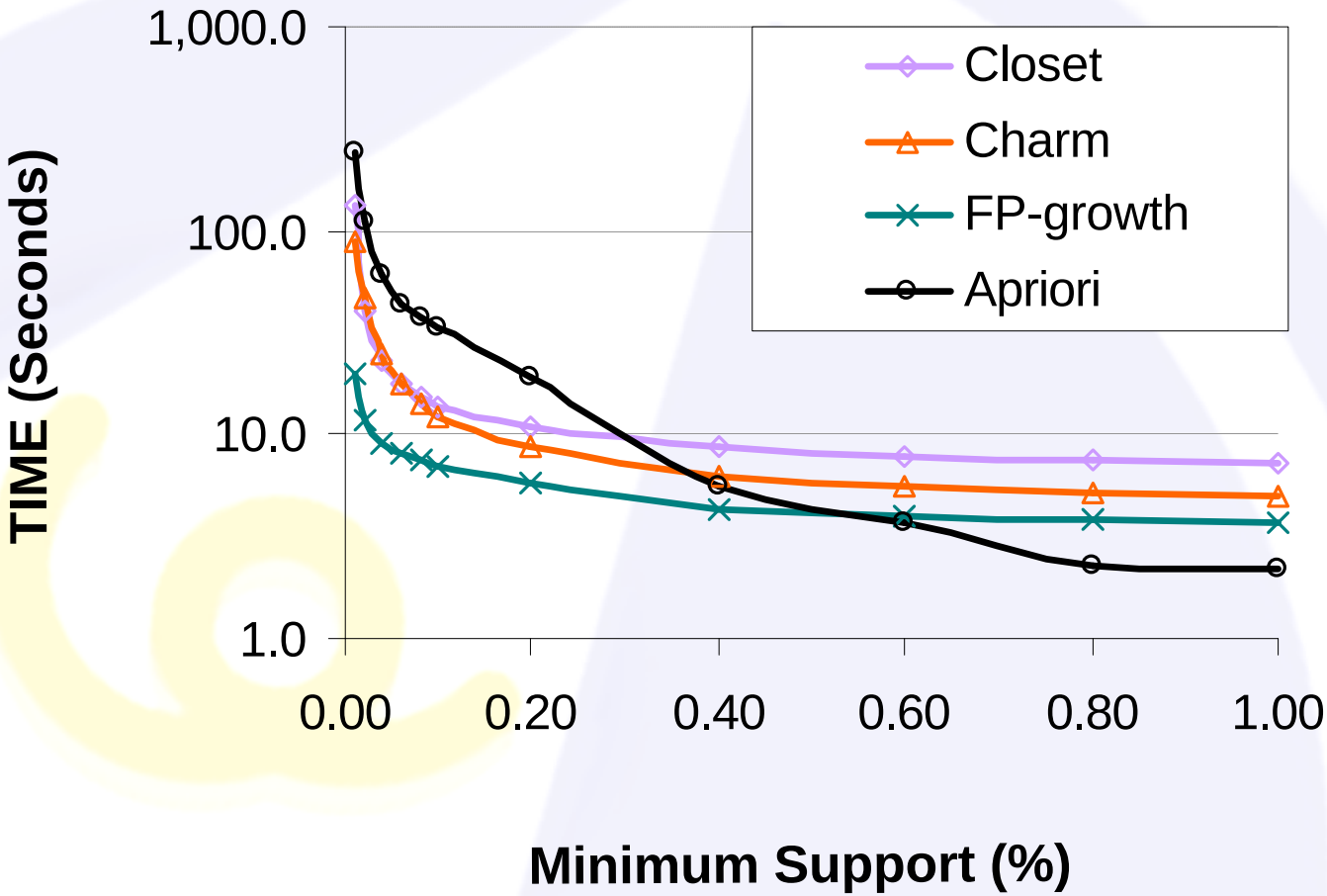
Ranking

	High Min-Support	**Low Min-Support**
IBM-Artificial	Ap > FP > Ch > Cl	FP > Ch > Cl > Ap
BMS-POS	Ap > Cl > FP > Ch	Ch > FP > Ap > Cl
BMS-WebView-1	Ap > FP > Cl > Ch	Ch > FP > Ap > Cl
BMS-WebView-2	Ap > FP > Ch > Cl	Ch > FP > Ap > Cl

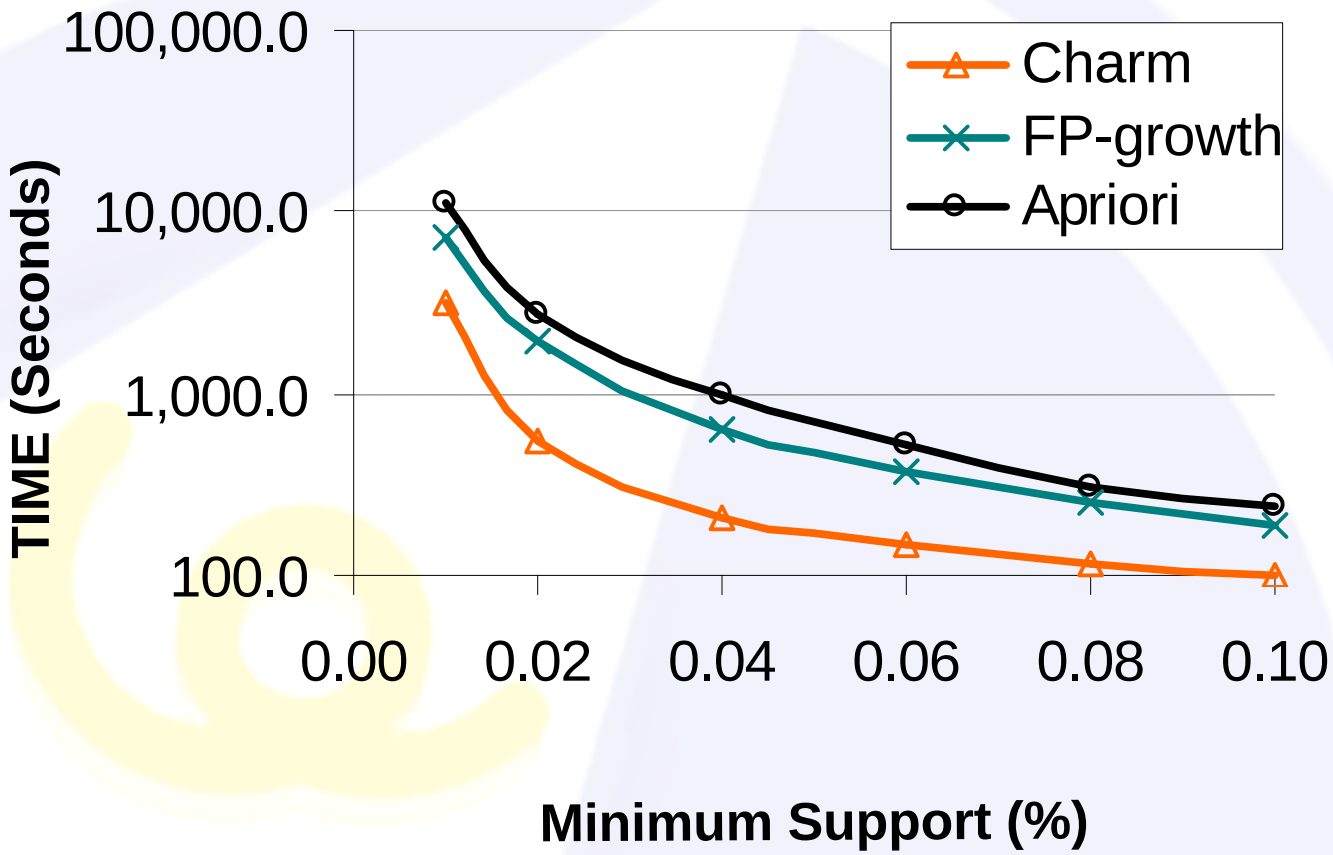
IBM-Artificial (frequent itemsets)



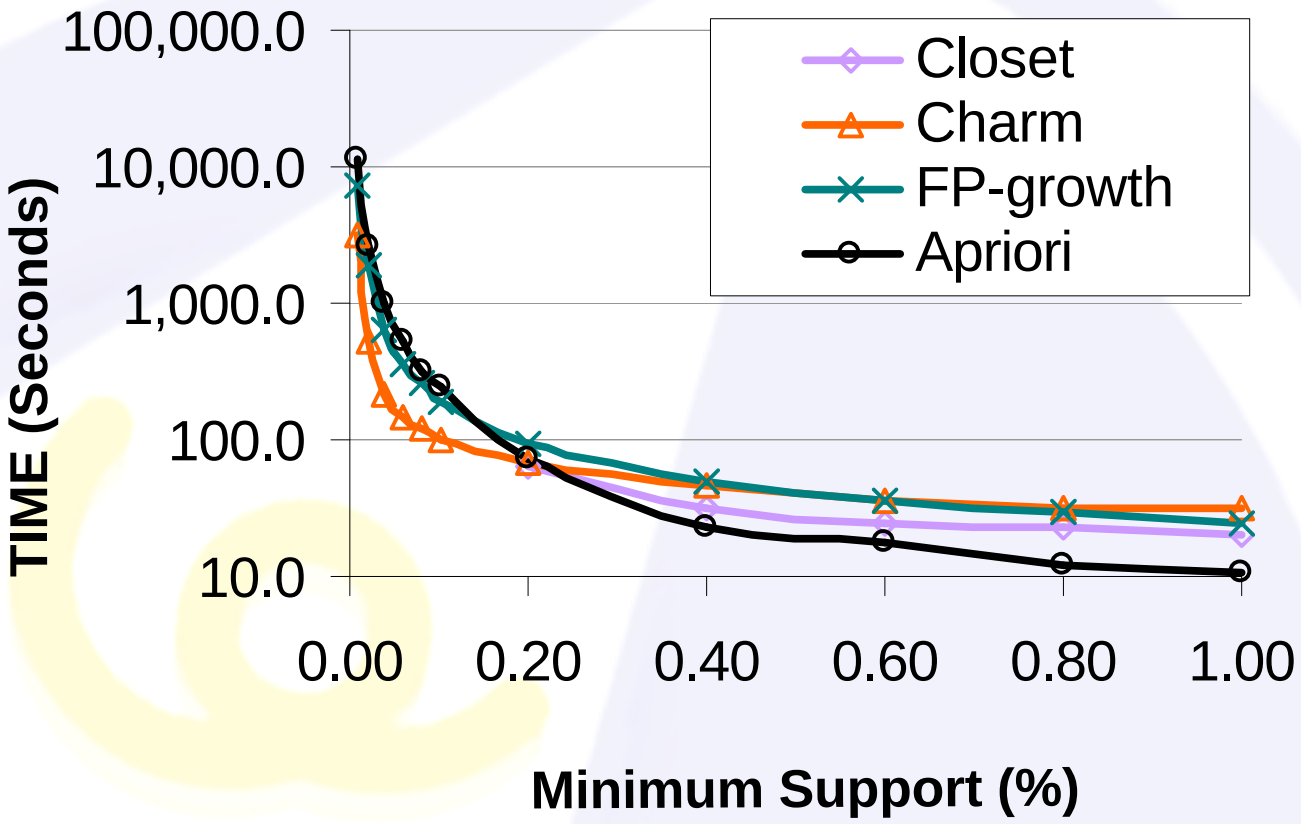
IBM-Artificial (frequent itemsets)



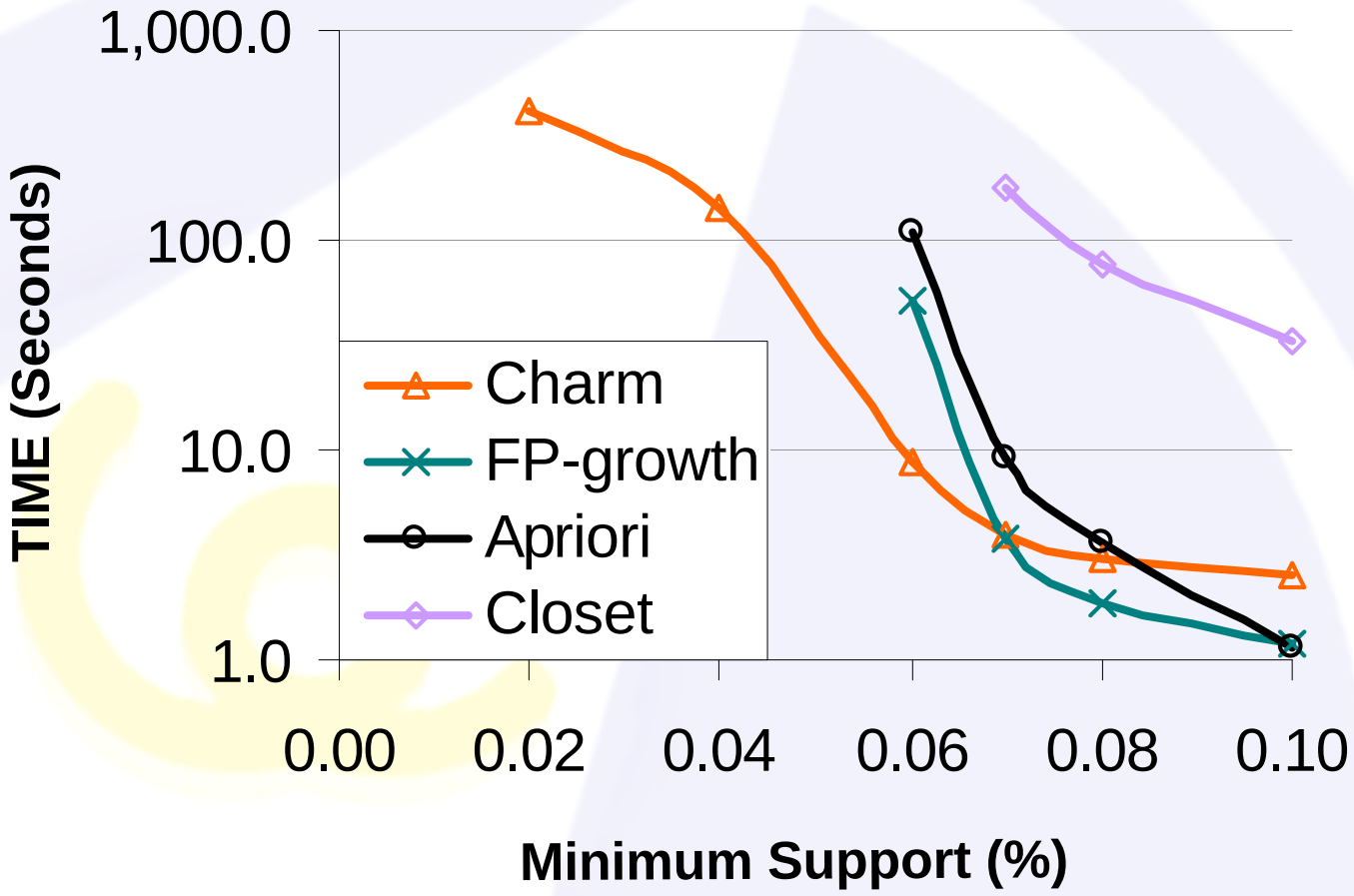
BMS-POS (frequent itemsets)



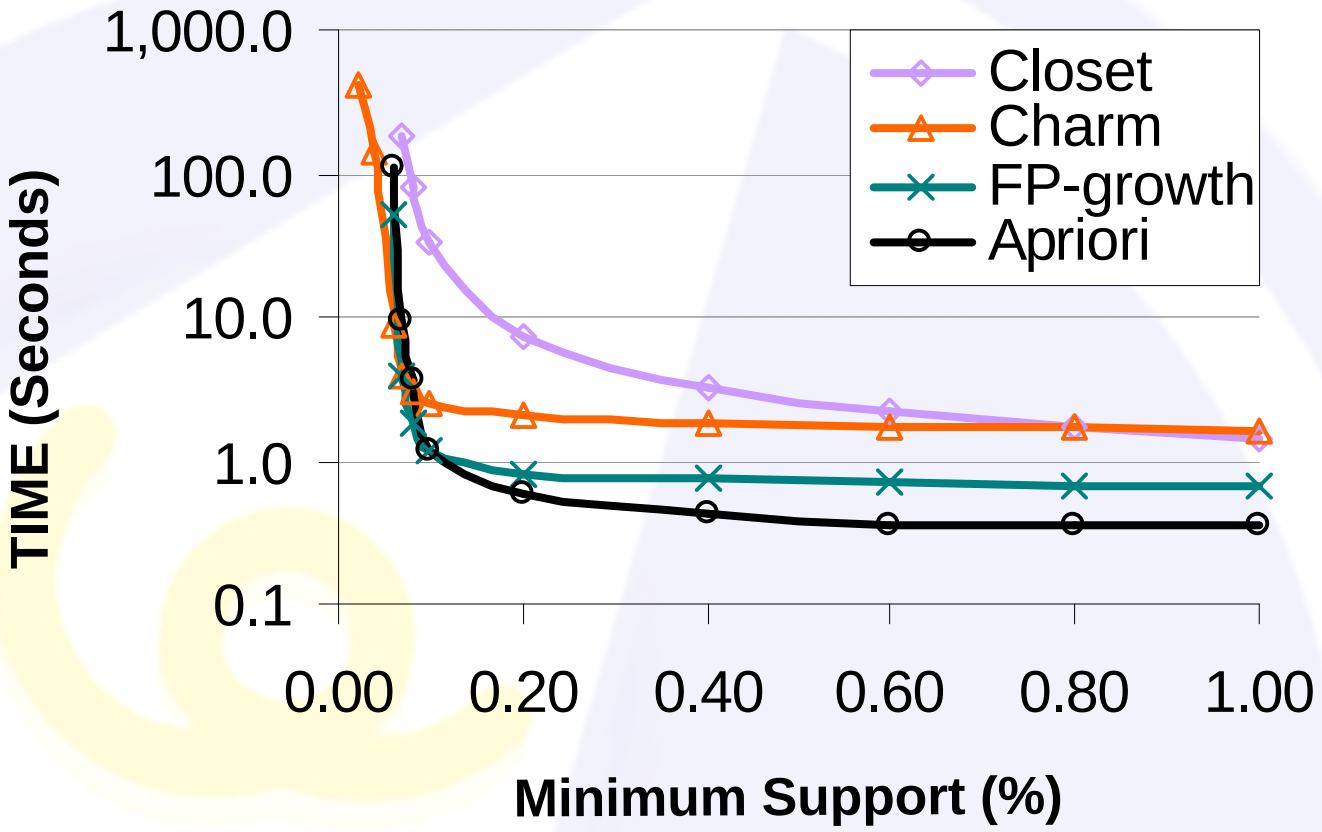
BMS-POS (frequent itemsets)



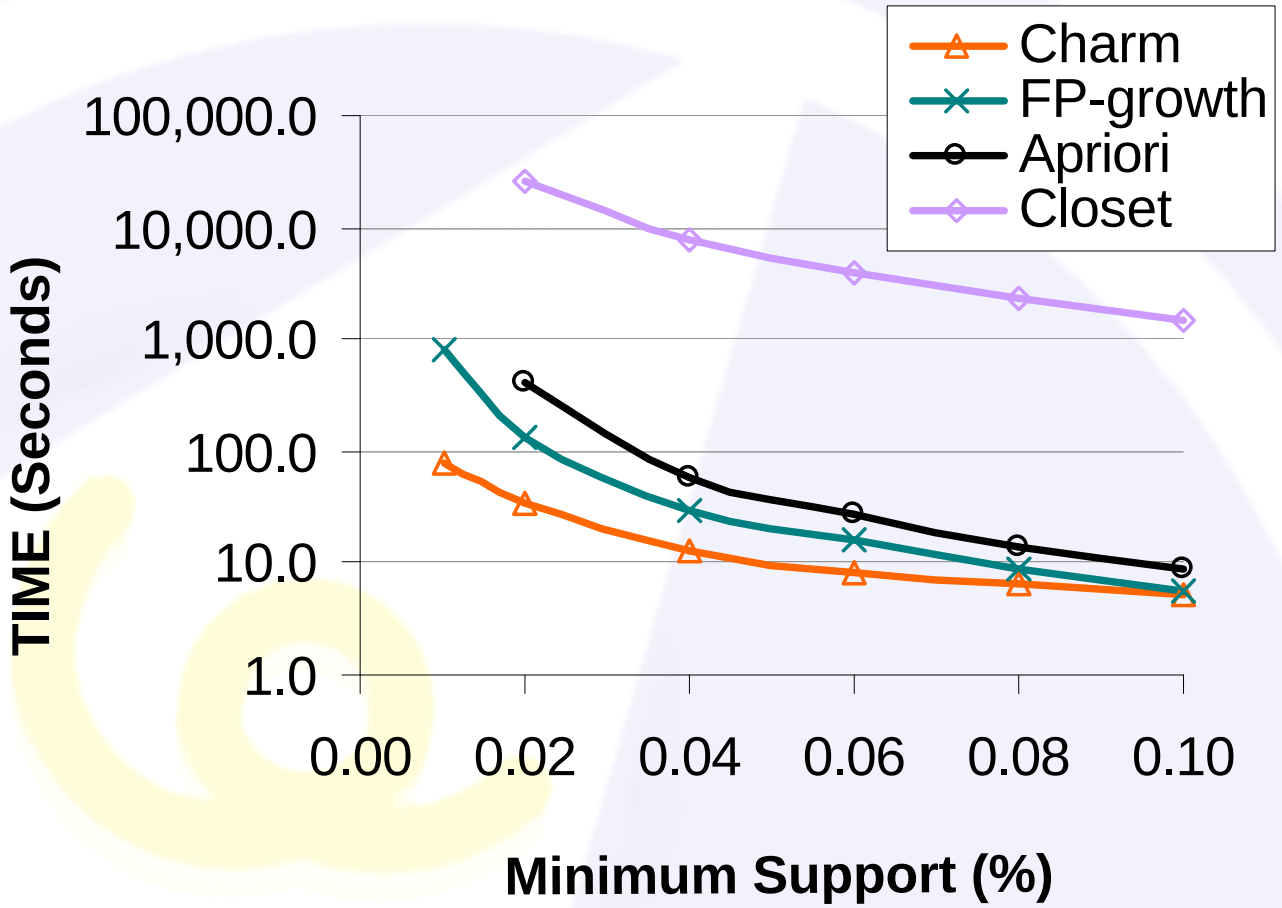
BMS-WebView-1 (frequent itemsets)



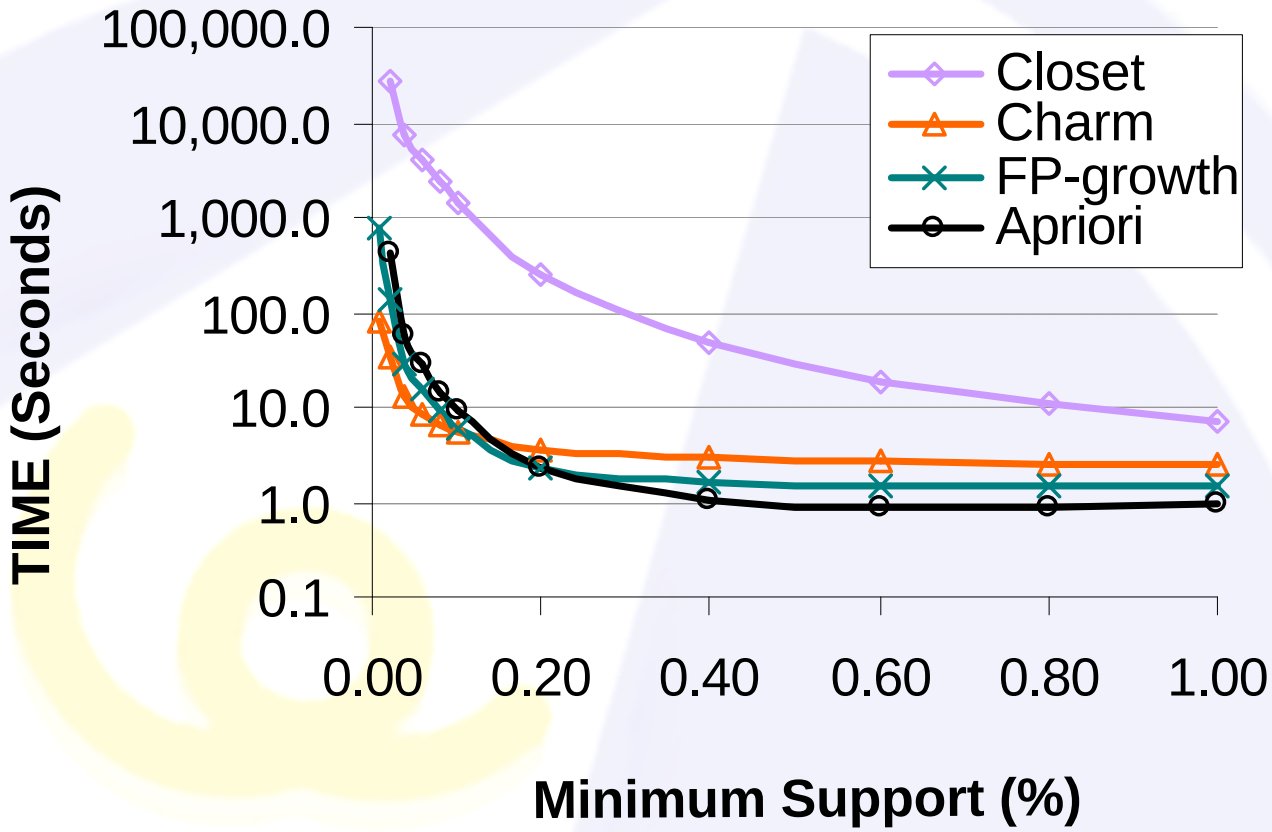
BMS-WebView-1 (frequent itemsets)



BMS-WebView-2 (frequent itemsets)



BMS-WebView-2 (frequent itemsets)



Experimental Results (Cont'd)

BLUE MARTINI SOFTWARE

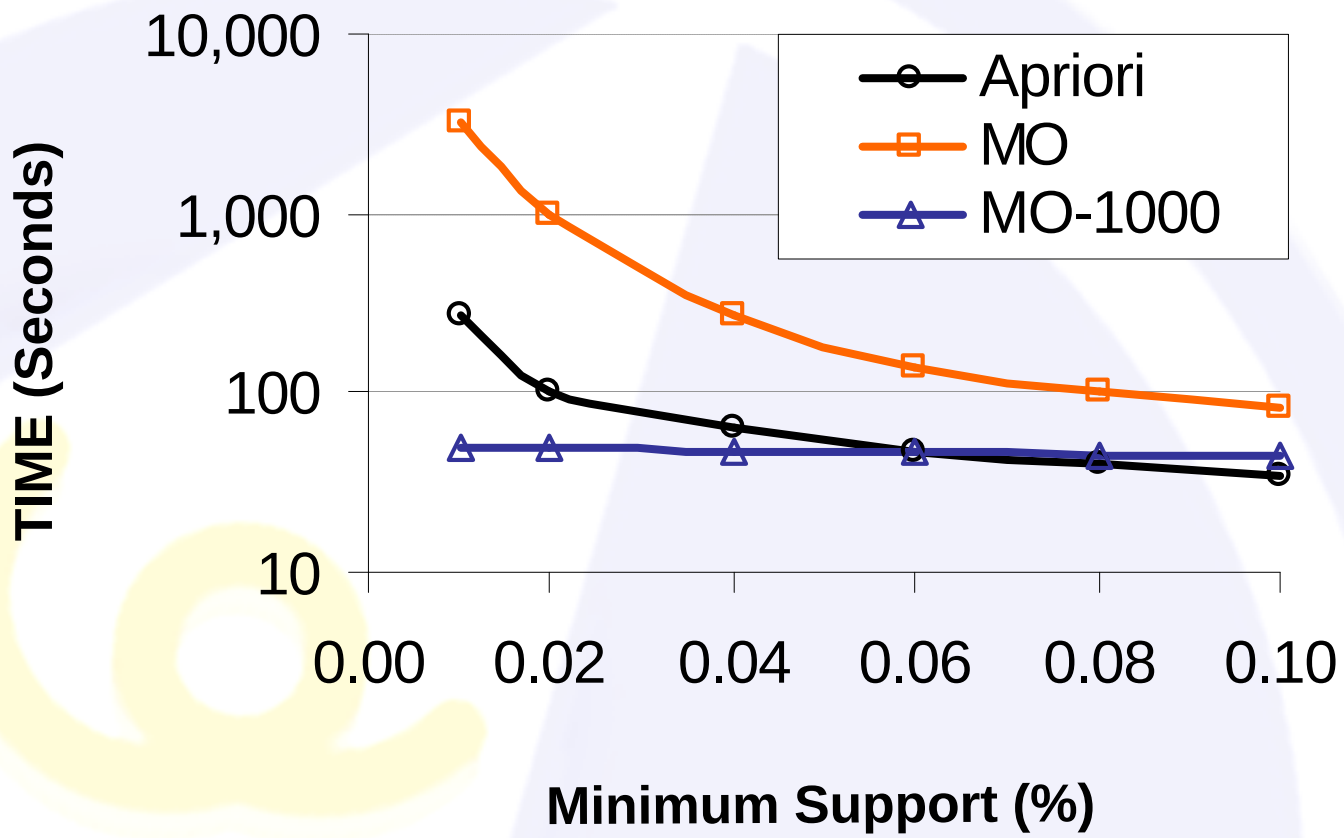
On some real-world datasets, when the minimum support is small, the number of frequent itemsets increases super-exponentially, thus no algorithm can handle it.

E.g. BMS-WebView-1:

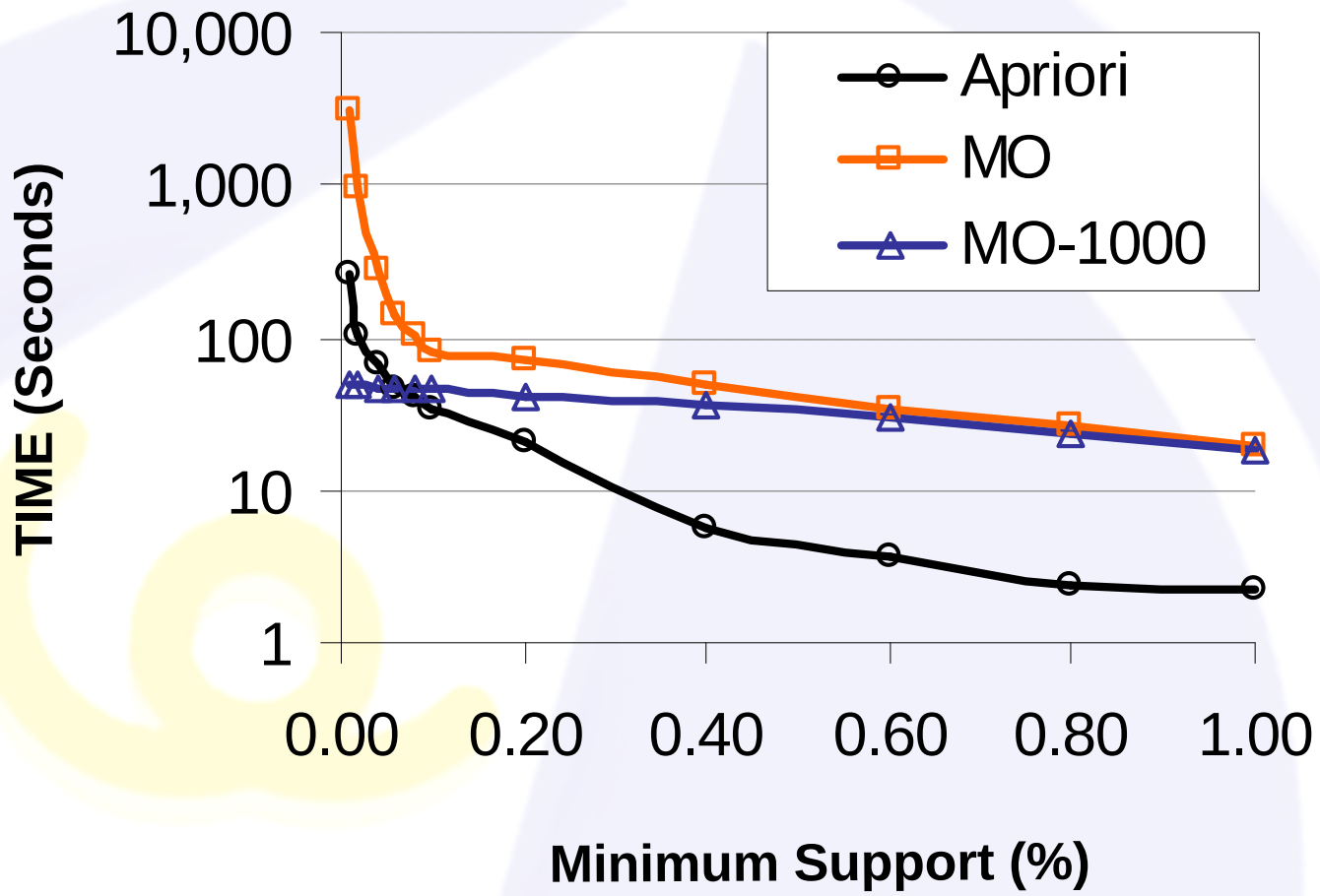
Minimum support(%)	Frequent itemsets
0.06	461,521
0.04	* 6.82×10^{10}
0.02	* 1.08×10^{26}
0.01	* 1.78×10^{45}

*: estimated

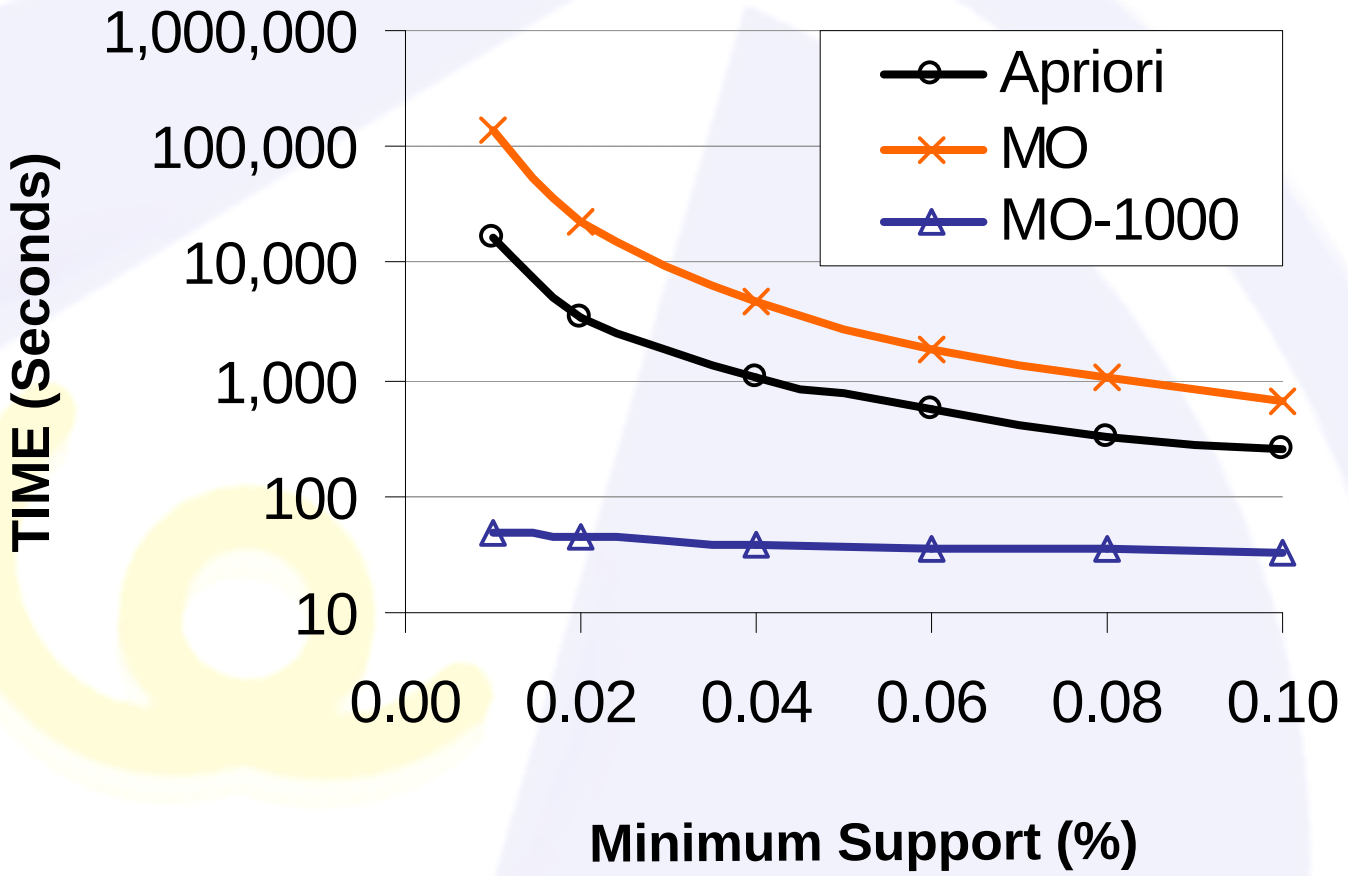
IBM-Artificial (association rules)



IBM-Artificial (association rules)

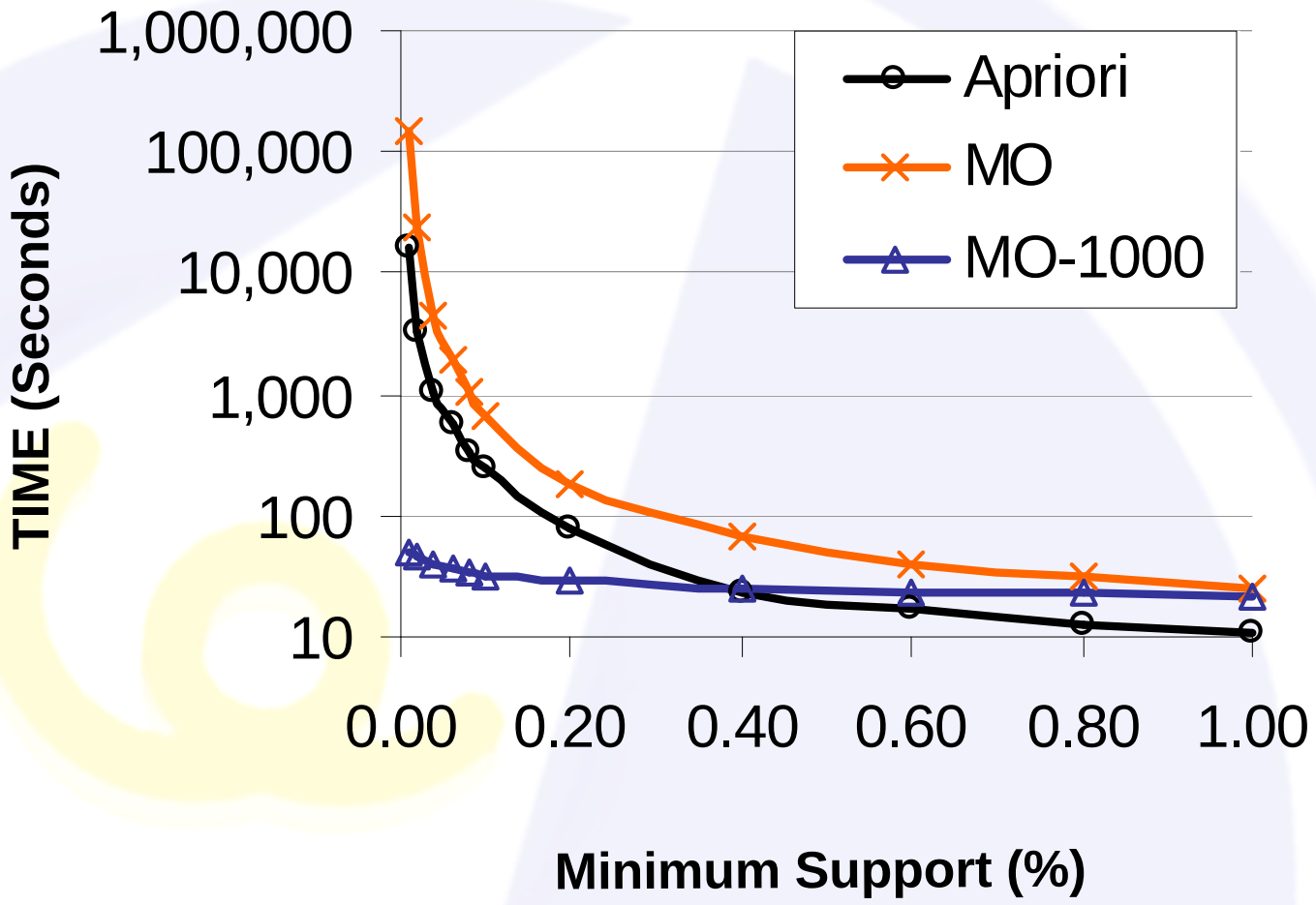


BMS-POS (association rules)



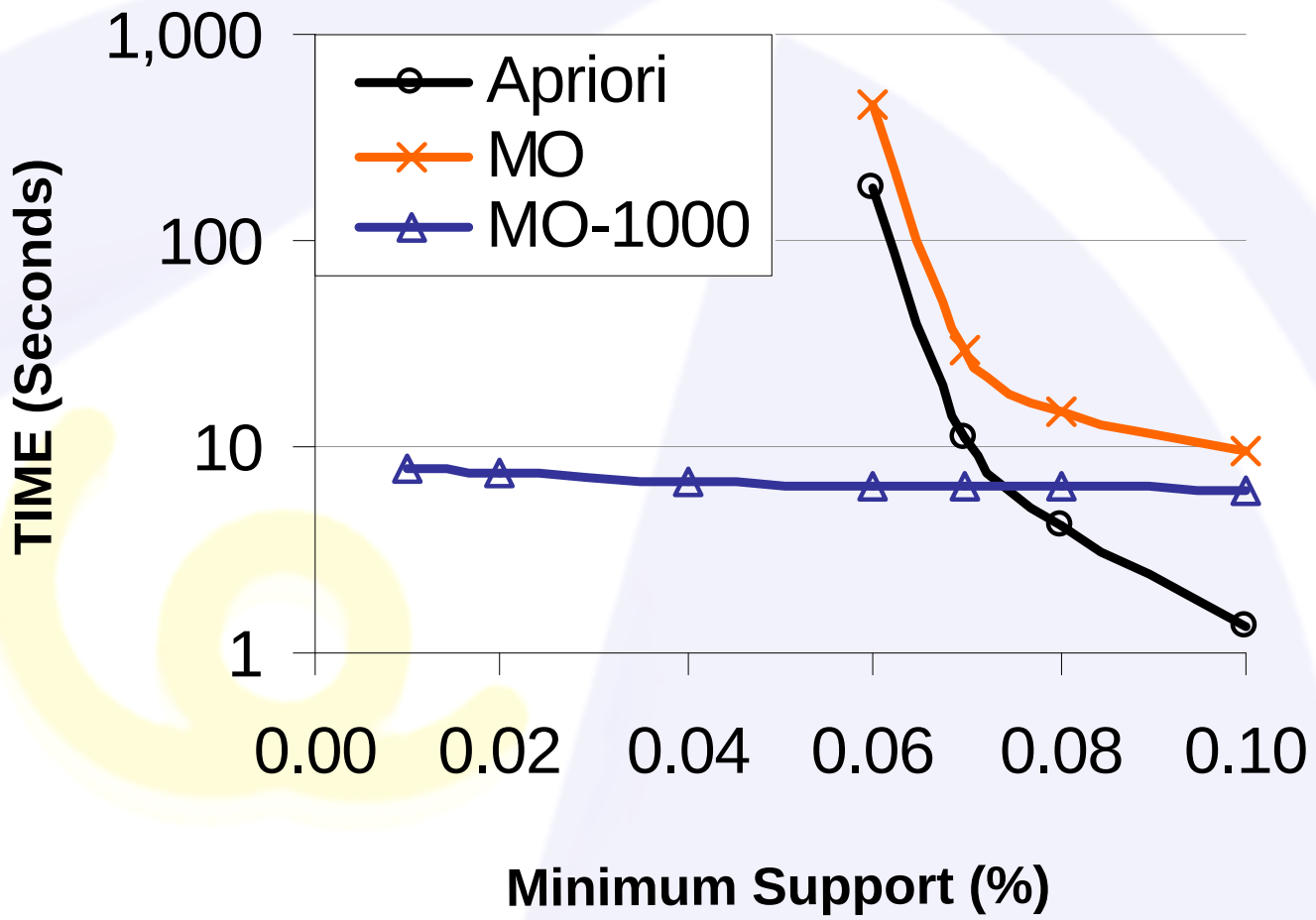
BMS-POS (association rules)

BLUE MARTINI SOFTWARE

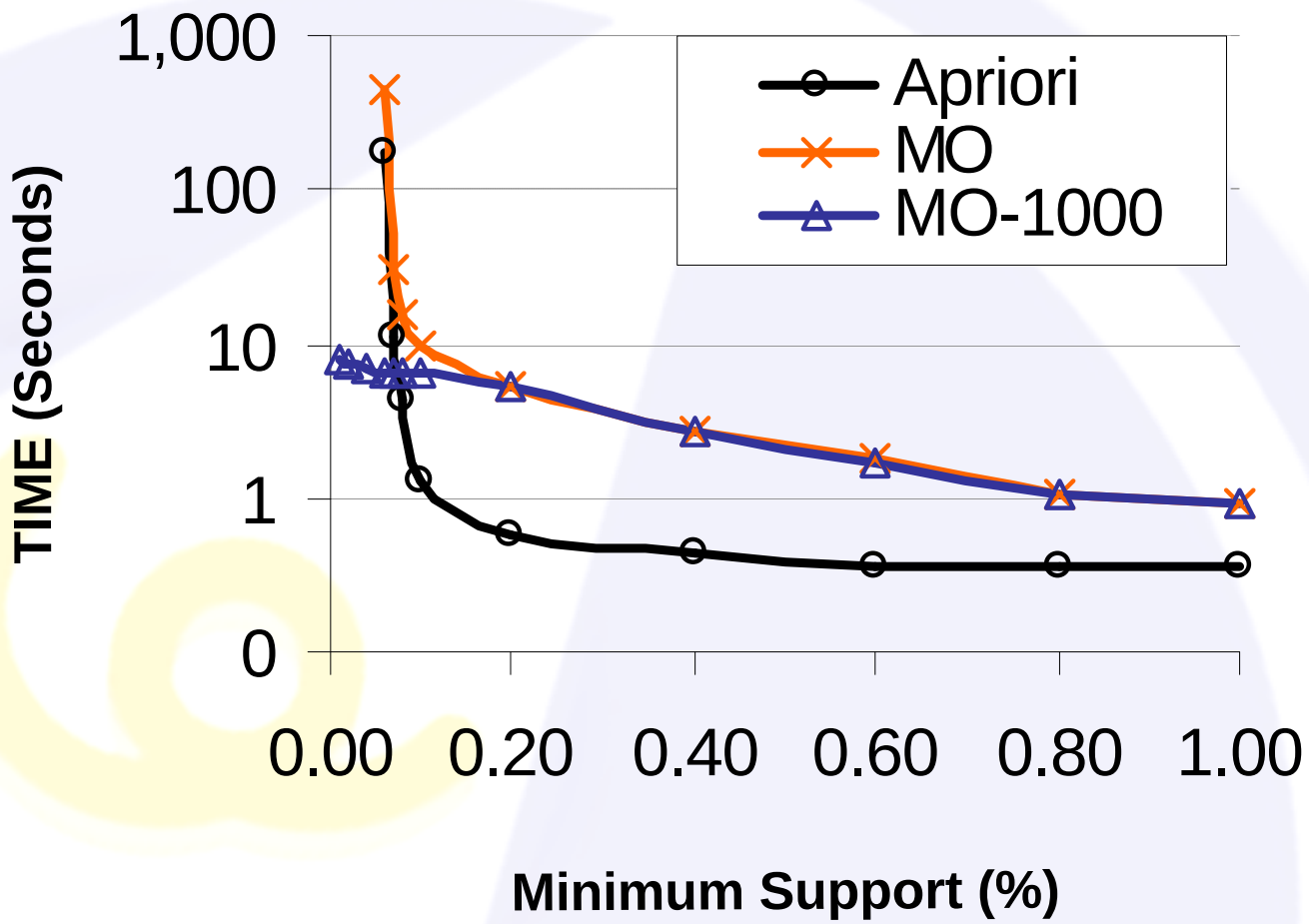


BMS-WebView-1 (association rules)

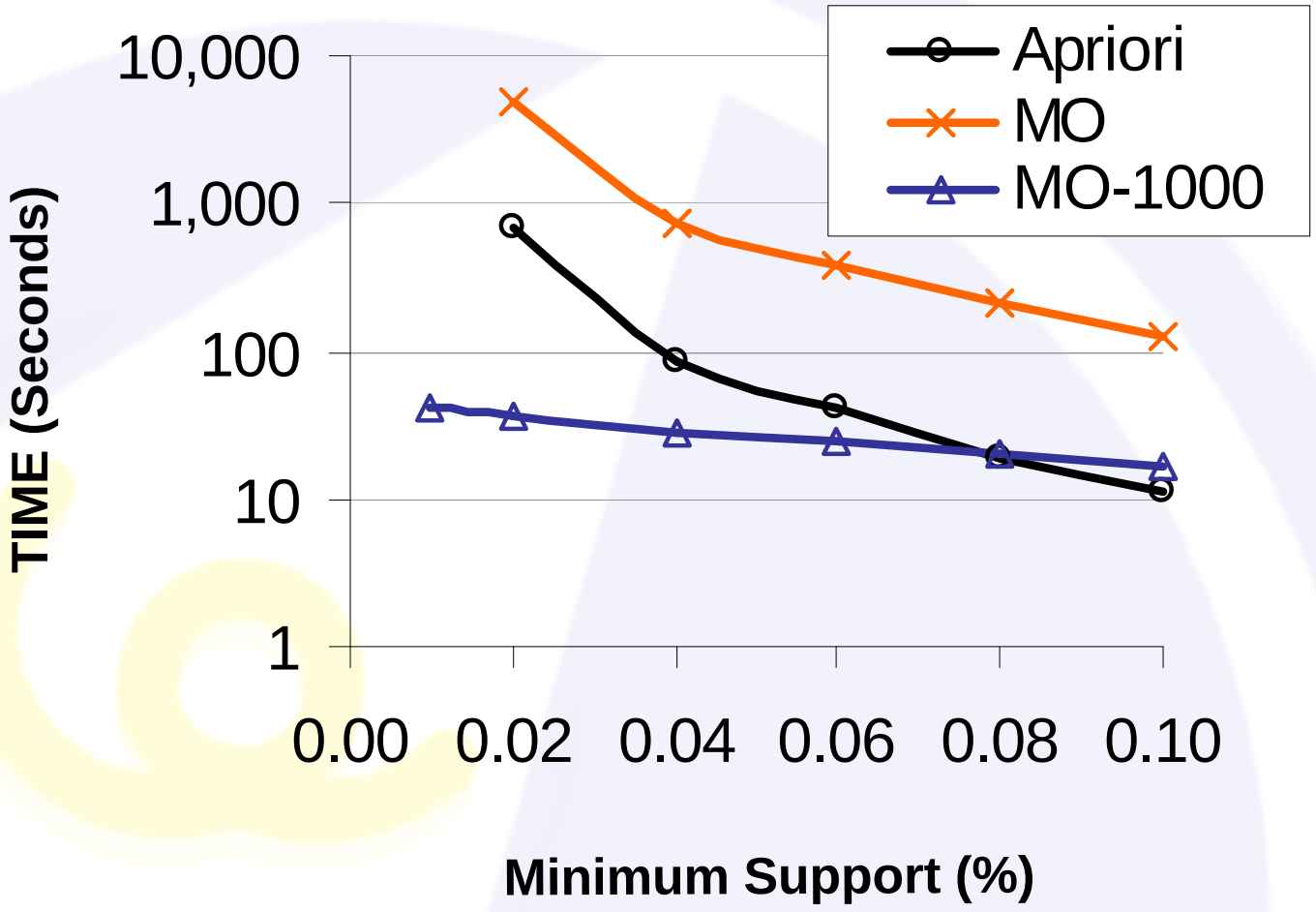
BLUE MARTINI SOFTWARE



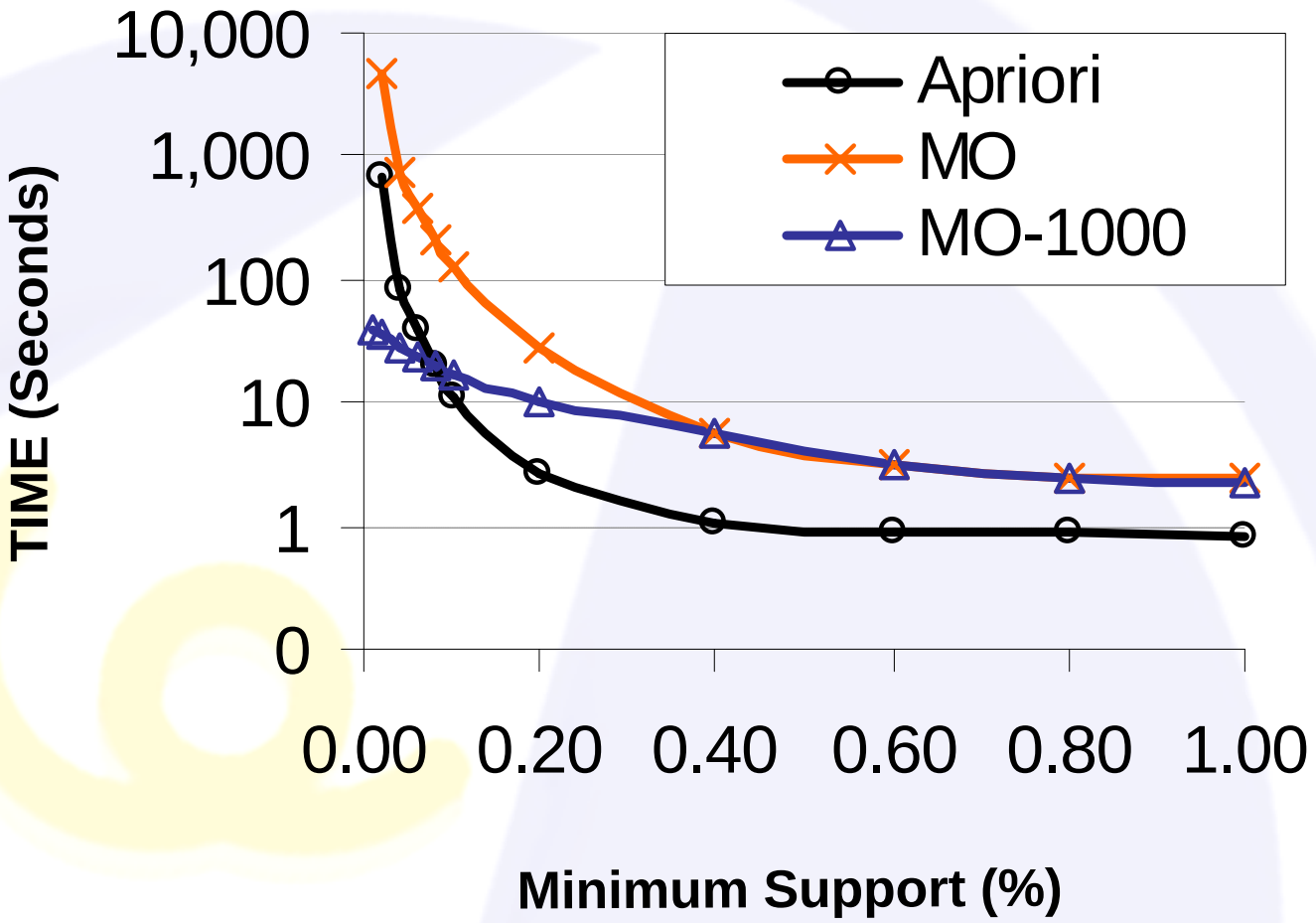
BMS-WebView-1 (association rules)



BMS-WebView-2 (association rules)



BMS-WebView-2 (association rules)



Experimental Results (Cont'd)

BLUE MARTINI SOFTWARE

- **MO could be a solution when**
  - **the number of association rules is very large**
  - **only the top-N rules are needed (based on some criteria such as lift or confidence)**

Contributions

- First objective evaluation and comparison of association rule algorithms on real-world e-commerce and retail datasets.
- Donated one e-commerce datasets for use in the research community.

Contributions (Cont'd)

BLUE MARTINI SOFTWARE

- Artificial datasets have very different characteristics from the real-world datasets.
- On real world data:
  - Very narrow min-sup range of interest
  - Super exponential growth in number of rules
- Performance improvements on artificial data did not generalize to real world data
- Optimizing algorithms for these artificial datasets can mislead research effort