Mining Attribute-structure Correlated Patterns in Large Attributed Graphs

@VLDB'12, Istanbul, Turkey

Arlei Silva[1,3], Wagner Meira Jr.[1], Mohammed J. Zaki[2]

[1]Computer Science Department – Universidade Federal de Minas Gerais, Brasil

[2]Computer Science Department – Rensselaer Polytechnic Institute, USA

[3]Computer Science Department – University of California, Santa Barbara, USA

inweb

Motivation: Graph Analysis

- Powerful framework for the analysis of interaction data
- More data $\rightarrow$ more complex representations $\rightarrow$ new patterns
- Real graphs have not only structural but also vertex attribute information
- Existing graph mining techniques are not able to associate vertex attributes and dense subgraphs

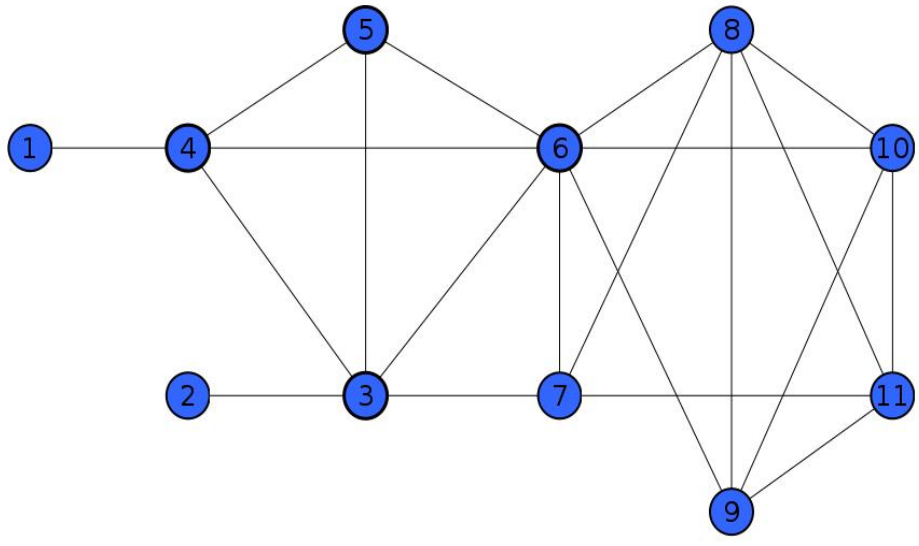
Attributed Graphs

- **Vertex attributes + graph structure**

- **Examples:**
  - Personal characteristics + social network
  - Content + citation structure

vertex	attributes
1	A, C
2	A, C
3	A, C, D
4	A, D
5	A, C, E
6	A, B, C
7	A, B, E
8	A, B
9	A, B
10	A, B, D
11	A, B, C

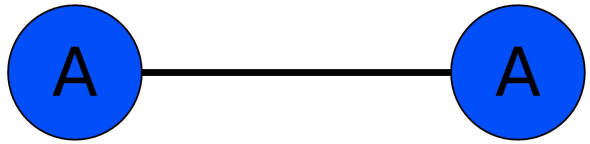
(a) Vertex attributes



(b) Graph

Structural Correlation in Attributed Graphs

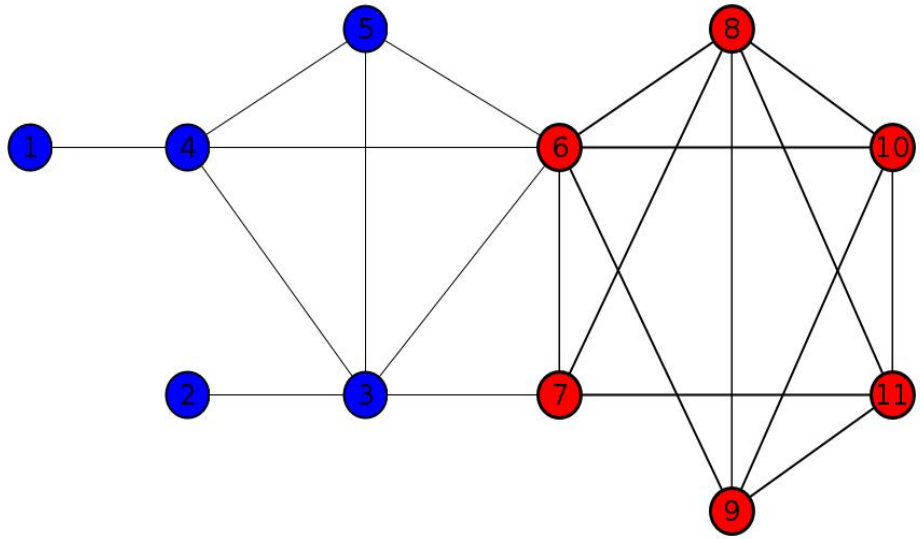
Two vertices correlated wrt an attribute A:



Structural correlations = attributes X dense subgraphs

- Densely connected webpages that share content
- Groups of friends with common interests
- Genes that interact and are expressed on the same tissues

vertex	attributes
1	A, C
2	A, C
3	A, C, D
4	A, D
5	A, C, E
6	A, B, C
7	A, B, E
8	A, B
9	A, B
10	A, B, D
11	A, B, C



Related Work

Dense subgraphs [Gibson et al.'05, Pei et al.'05, Wang et al.'06, Zeng et al'07, Liu and Wong'08, Jiang and Pei'09] **and graph communities** [Girvan and Newman'02, Fortunato'10].

Frequent itemset mining[Agrawal et al.'93, Zaki'00].

Attributes + structure [Ge et al.'08, Zhou et al.'09, Khan et al.'10, Moser et al.'09, Mougel et al.'10, Sese et al.'10]

Structural correlation [Anagnostopoulos et al.'08, Silva et al.'10, Guan et al.'11, Wu et al.'12, Guan et al.'12, Prado et al.'12]

Structural Correlation Pattern Mining: Definitions

Dense subgraph (Q): Maximal vertex set Q such that for each $v \in Q$, the degree of v in Q is, at least, $\lceil \gamma_{min}.(|Q| - 1) \rceil$.

Structural correlation (ϵ): Probability of a vertex that has an attribute set S to be part of a correlated dense subgraph Q.

$$\epsilon(S) = \frac{|\mathcal{K}_S|}{|\mathcal{V}(S)|}$$

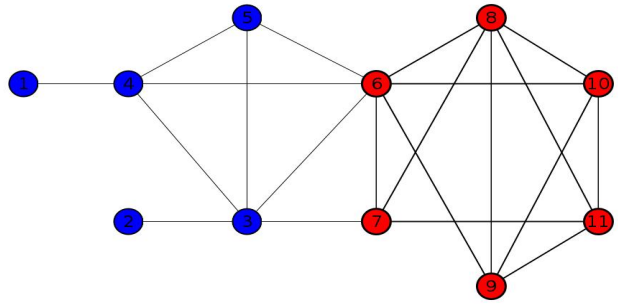
where $\mathcal{K}_S$ is the structural coverage of S.

Structural correlation pattern (S, Q): Correlated dense subgraph Q wrt S.

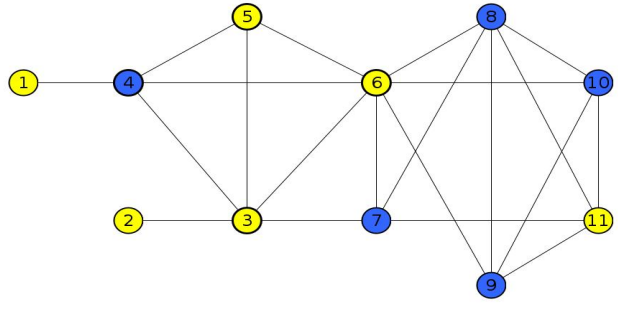
Structural Correlation Pattern Mining: Examples

vertex	attributes
1	A, C
2	A, C
3	A, C, D
4	A, D
5	A, C, E
6	A, B, C
7	A, B, E
8	A, B
9	A, B
10	A, B, D
11	A, B, C

(a) Vertex attributes



(b) $\epsilon(B) = 1.0$



(c) $\epsilon(C) = 0$

$(\{B\}, \{6, 7, 8, 9, 10, 11\})$ is a structural correlation pattern

Structural Correlation Pattern Mining: Challenges

SCP mining is #P-hard

- Quasi-clique mining (#P-hard) $\overset{poly}{\rightarrow}$ SCP mining
  1. Pick up an arbitrary non-attributed graph $\mathcal{G}$
  2. Make $\mathcal{G}$ attributed by adding a single attribute to all vertices
  3. Set σ=1 and ϵ_{min}=0

Support bias

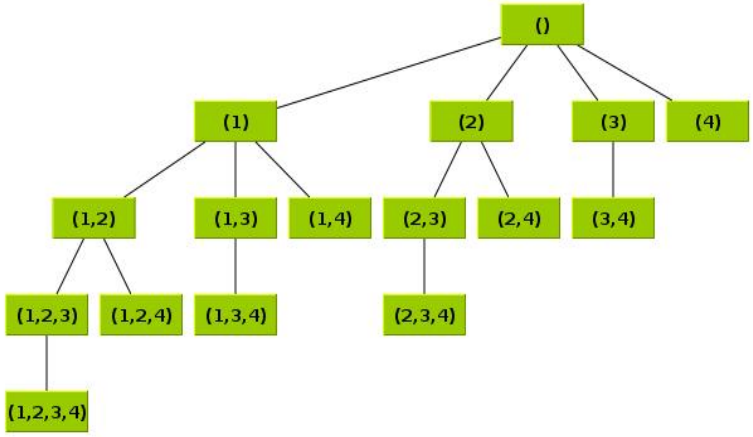
- Expected structural correlation $\propto$ attribute set frequency
  - Frequent attribute sets are more likely to be shared by vertices that are part of a dense subgraph

SCPM Algorithm

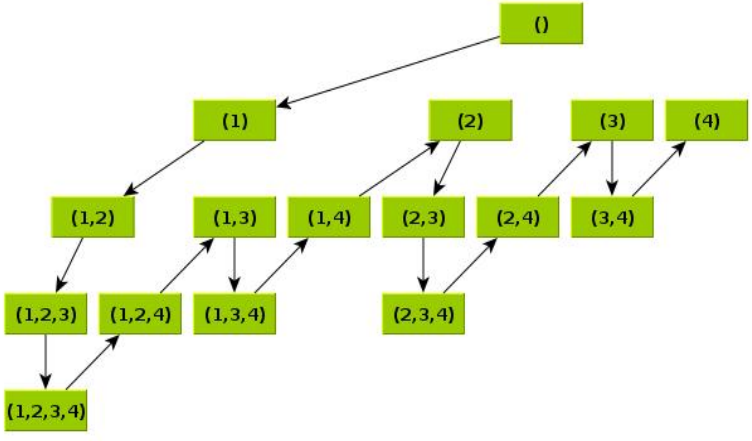
SCPM combines frequent itemset and quasi-clique mining strategies with effective search, pruning, and significance measures for structural correlation pattern mining.

- **Search (DFS or BFS)**
- **Pruning (Vertex + attribute sets)**
- **Statistical significance of correlations**
- Top-k patterns

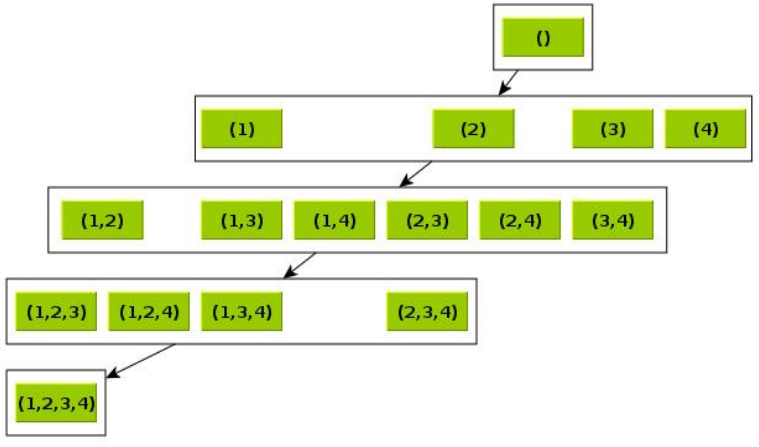
SCPM: Searching for Patterns



(a) Search space



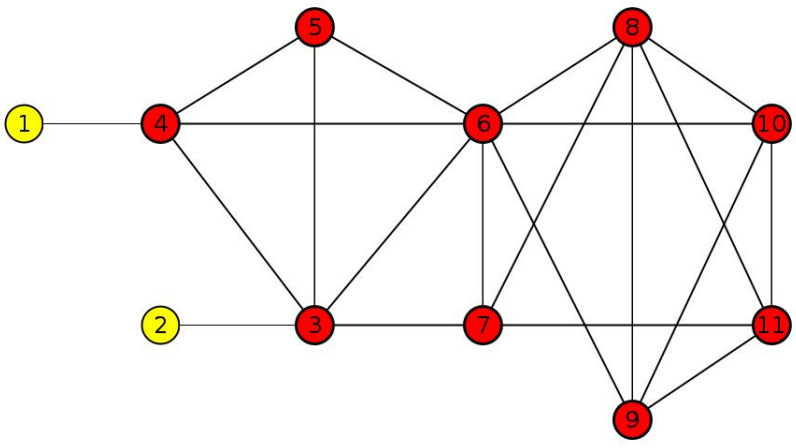
(b) DFS



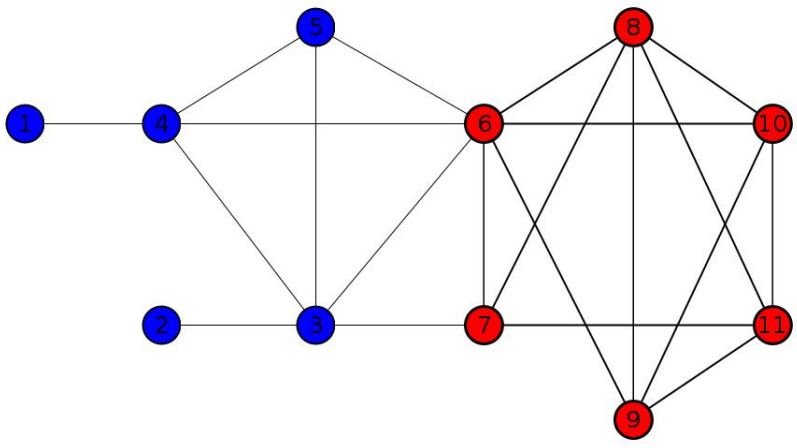
(c) BFS

SCPM: Vertex Pruning

Vertex pruning: If S_i is a subset of S_j, then vertices covered by dense subgraphs in $\mathcal{G}(S_j)$ must also be covered in $\mathcal{G}(S_i)$.



(a) $\mathcal{G}(A)$



(b) $\mathcal{G}(A, B)$

SCPM: Attribute Set Pruning

Attribute set pruning: **If S_i is a subset of S_j and $\sigma(S_j) \geq \sigma_{min}$, then $\epsilon(S_j)$ is upper bounded by $\epsilon(S_i).|\mathcal{V}(S_i)|/\sigma_{min}$.**

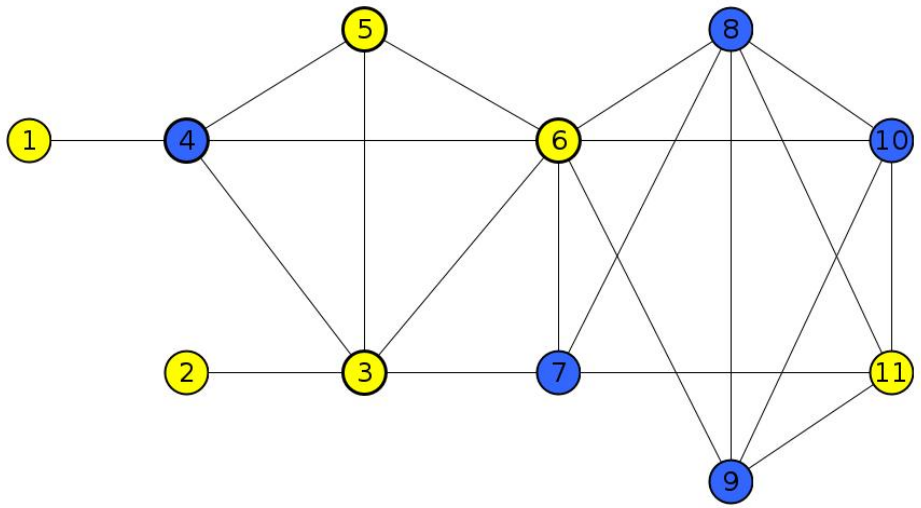
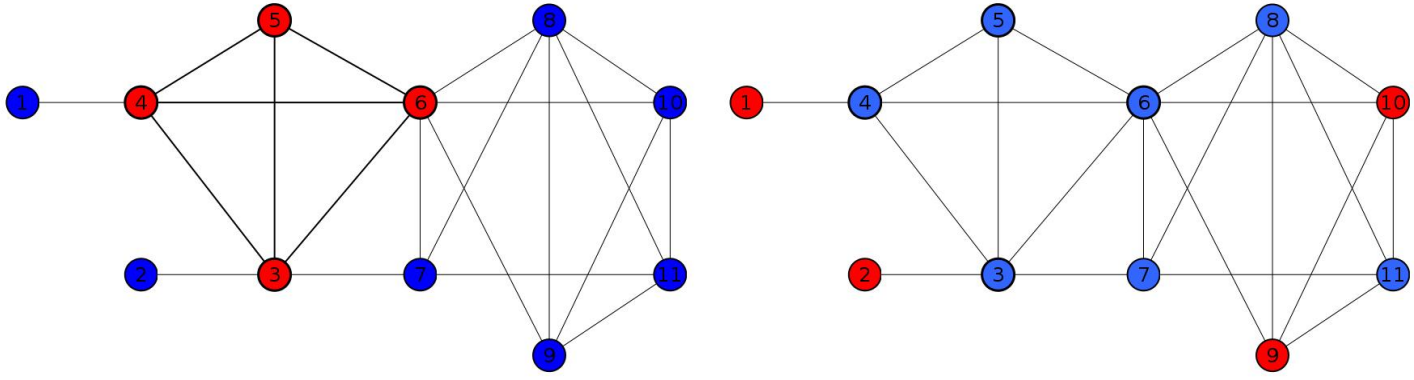


Figure: $\epsilon(C) = 0$. If $C \subseteq S_j$ then $\epsilon(S_j) = 0$

.

SCPM: Statistical Significance



(a) Lucky case (b) Unlucky case

330 possible random subgraphs of size 4

$$max\text{-}\epsilon_{exp}(\sigma, \gamma_{min}, min\_size) = \sum_{\alpha=z}^{m} p(\alpha). \sum_{\beta=z}^{\alpha} F(\alpha, \beta, \rho)$$

where:

- $z = \lceil \gamma_{min}.(min\_size - 1) \rceil$
- m is the max degree in $\mathcal{G}$
- p is the degree distribution of $\mathcal{G}$
- $F(\alpha, \beta, \rho) = \binom{\alpha}{\beta}.\rho^{\beta}.(1-\rho)^{\alpha-\beta}$
- $\rho = \frac{\sigma-1}{|\mathcal{V}-1|}$

Experimental Results: Datasets

name	vertex	edge	attribute	attr. set	subgraph
DBLP	author	co-authorship	term	topic	community
Citeseer	paper	citation	term	topic	related work
LastFm	user	friendship	artist	taste	community

Table: Descriptions

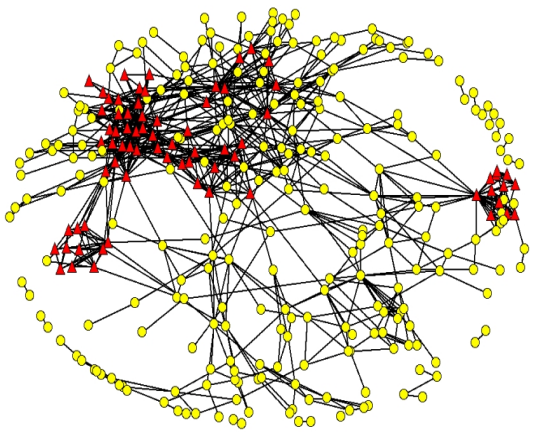
name	#vertices	#edges	#attributes
DBLP	108,030	276,658	23,285
Citeseer	294,104	782,147	206,430
LastFm	272,412	350,239	3,929,101

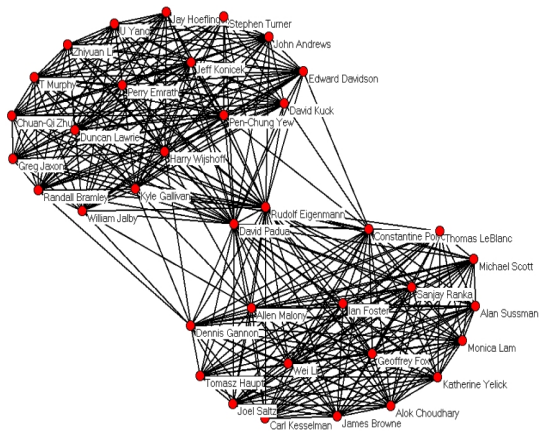
Table: Statistics

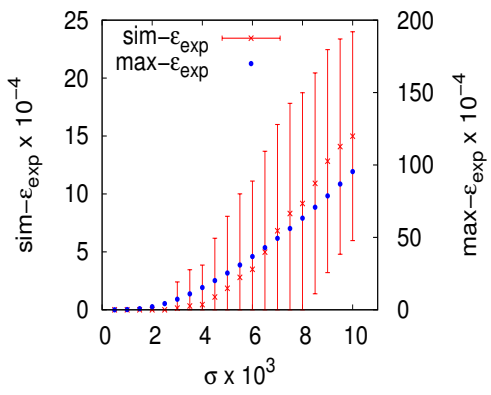
Experimental Results: DBLP

top-σ (support)				top-ϵ (structural correlation)				top-δ_{lb} (norm. str. correlation)			
S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}
base system	5,492	.04	14.0	grid applic	840	.26	41,577	search rank	420	.19	635,349
base us	5,421	.04	13.5	grid servic	599	.23	154,703	perform file	404	.14	555,067
base model	4,852	.03	13.3	environ grid	525	.21	256,793	structur index	404	.14	555,067
model us	4,168	.03	21.0	queri xml	615	.21	123,533	search mine	413	.14	490,932
system us	3,989	.05	36.8	search web	1,031	.20	13,738	us xml	400	.11	442,638

Table: Attribute sets







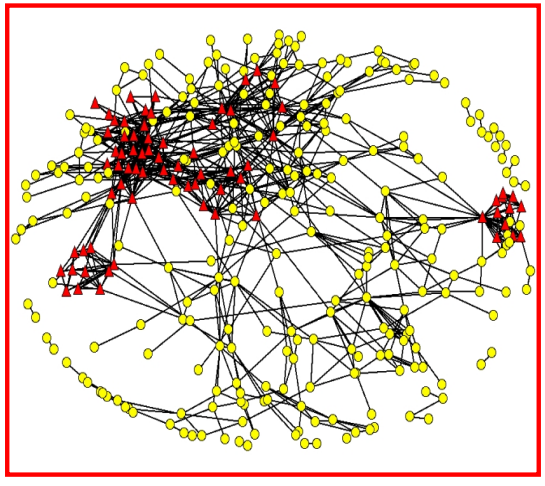
(a) $\mathcal{G}(\{search, rank\})$ (b) $\{system, performance\}$ (c) $\sigma \times \epsilon_{exp}$

$$\sigma_{min} = 400,\ |S|_{min} = 2,\ |Q|_{min} = 10,\ \gamma_{min} = 0.5$$

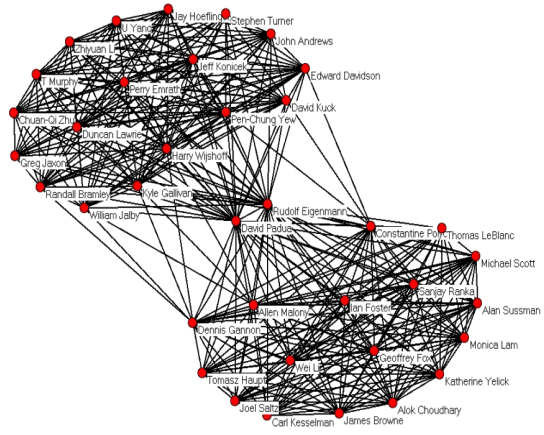
Experimental Results: DBLP

top-σ (support)				top-ϵ (structural correlation)				top-δ_{lb} (norm. str. correlation)			
S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}
base system	5,492	.04	14.0	grid applic	840	.26	41,577	**search rank**	**420**	**.19**	**635,349**
base us	5,421	.04	13.5	grid servic	599	.23	154,703	perform file	404	.14	555,067
base model	4,852	.03	13.3	environ grid	525	.21	256,793	structur index	404	.14	555,067
model us	4,168	.03	21.0	queri xml	615	.21	123,533	search mine	413	.14	490,932
system us	3,989	.05	36.8	search web	1,031	.20	13,738	us xml	400	.11	442,638

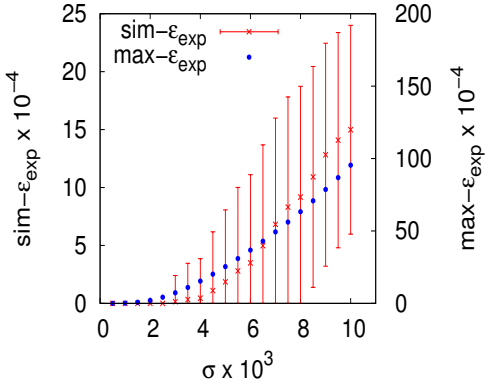
Table: Attribute sets



(a) $\mathcal{G}(\{search, rank\})$



(b) $\{system, performance\}$



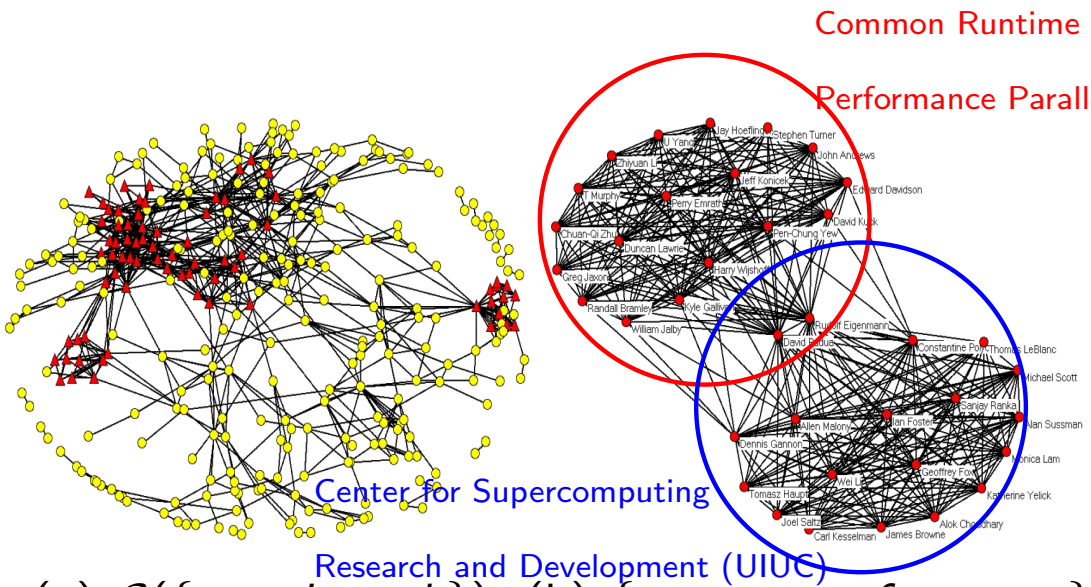
(c) $\sigma \times \epsilon_{exp}$

$$\sigma_{min} = 400,\ |S|_{min} = 2,\ |Q|_{min} = 10,\ \gamma_{min} = 0.5$$

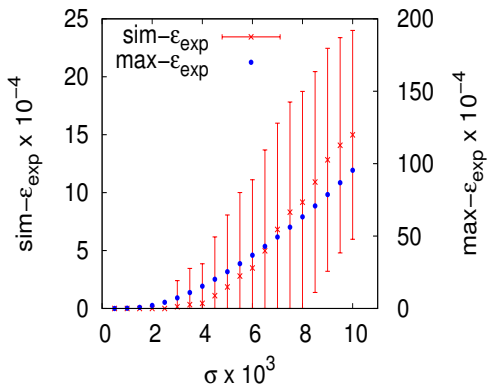
Experimental Results: DBLP

top-σ (support)				top-ϵ (structural correlation)				top-δ_{lb} (norm. str. correlation)			
S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}
base system	5,492	.04	14.0	grid applic	840	.26	41,577	search rank	420	.19	635,349
base us	5,421	.04	13.5	grid servic	599	.23	154,703	perform file	404	.14	555,067
base model	4,852	.03	13.3	environ grid	525	.21	256,793	structur index	404	.14	555,067
model us	4,168	.03	21.0	queri xml	615	.21	123,533	search mine	413	.14	490,932
system us	3,989	.05	36.8	search web	1,031	.20	13,738	us xml	400	.11	442,638

Table: Attribute sets



(a) $\mathcal{G}(\{search, rank\})$ (b) $\{system, performance\}$



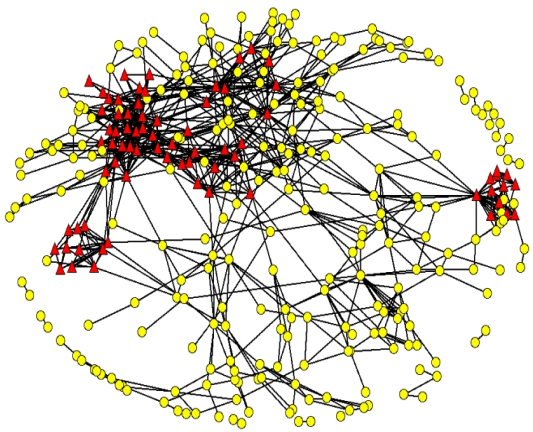
(c) $\sigma \times \epsilon_{exp}$

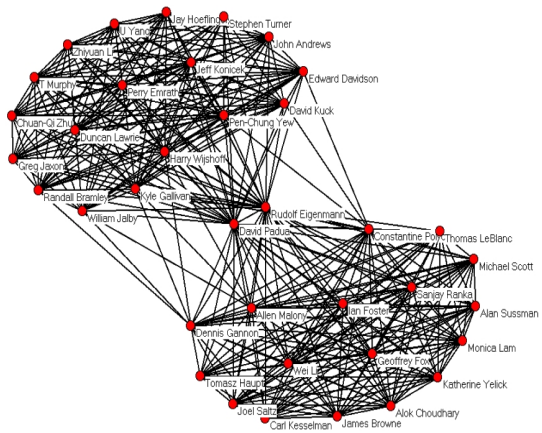
$$\sigma_{min} = 400,\ |S|_{min} = 2,\ |Q|_{min} = 10,\ \gamma_{min} = 0.5$$

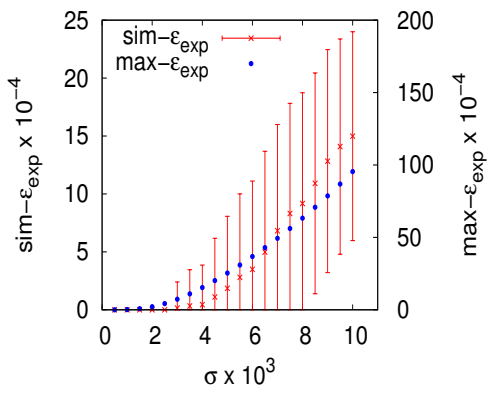
Experimental Results: DBLP

top-σ (support)				top-ϵ (structural correlation)				top-δ_{lb} (norm. str. correlation)			
S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}
base system	5,492	.04	14.0	grid applic	840	.26	41,577	search rank	420	.19	635,349
base us	5,421	.04	13.5	grid servic	599	.23	154,703	perform file	404	.14	555,067
base model	4,852	.03	13.3	environ grid	525	.21	256,793	structur index	404	.14	555,067
model us	4,168	.03	21.0	queri xml	615	.21	123,533	search mine	413	.14	490,932
system us	3,989	.05	36.8	search web	1,031	.20	13,738	us xml	400	.11	442,638

Table: Attribute sets







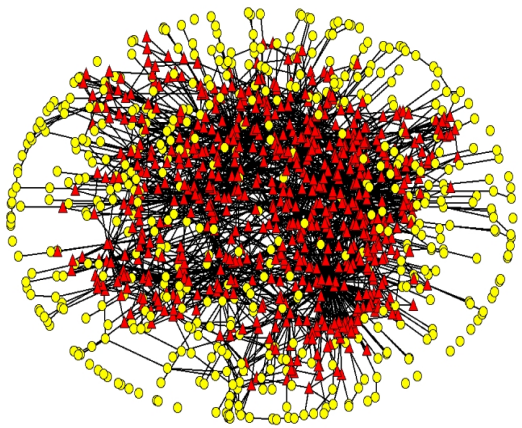
(a) $\mathcal{G}(\{search, rank\})$ (b) $\{system, performance\}$ (c) $\sigma \times \epsilon_{exp}$

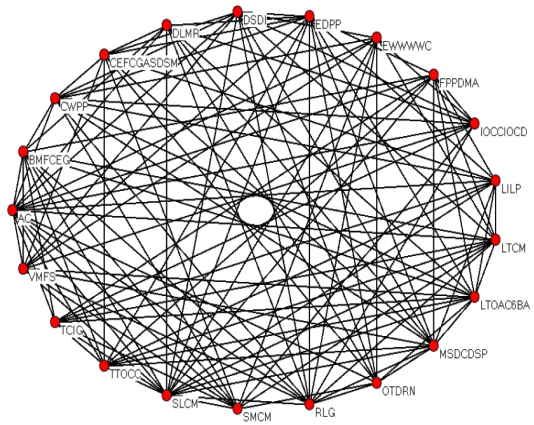
$$\sigma_{min} = 400,\ |S|_{min} = 2,\ |Q|_{min} = 10,\ \gamma_{min} = 0.5$$

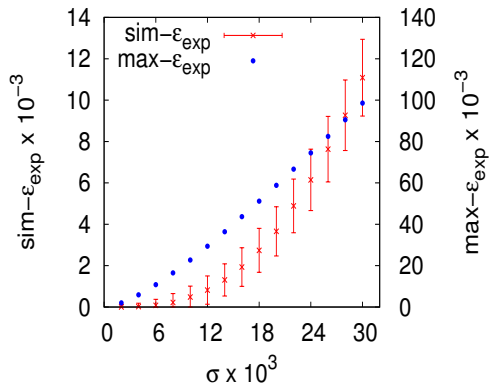
Experimental Results: CiteSeer

top-σ (support)				top-ϵ (structural correlation)				top-δ_{lb} (norm. str. correlation)			
S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}
system paper	57,906	.16	.77	network sensor	3,276	.47	108.7	node wireless	2,086	.35	164.4
base paper	56,566	.10	.47	network hoc	2,744	.47	141.2	protocol rout	2,134	.35	157.6
paper result	47,516	.08	.45	ad network hoc	2,725	.44	134.6	memori cach	2,150	.32	143.8
paper model	43,929	.09	.59	network rout	5,084	.41	48.0	network hoc	2,744	.47	141.2
us paper	43,573	.05	.32	network wireless	5,242	.40	44.7	protocol wireless	2,048	.29	138.7

Table: Attribute sets







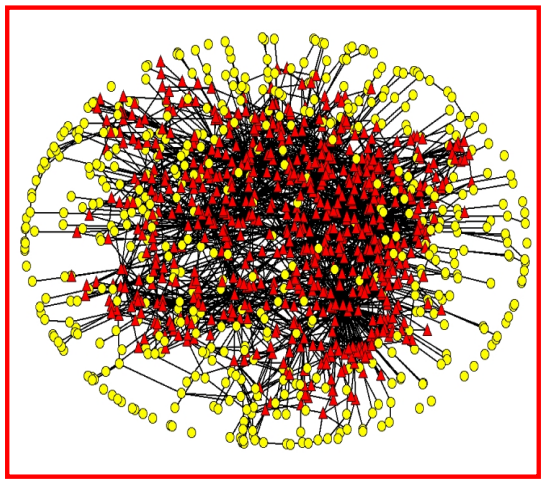
(a) $\mathcal{G}(\{node, wireless\})$ (b) $\{system, performance\}$ (c) $\sigma \times \epsilon_{exp}$

$$\sigma_{min} = 2,000,\ |S|_{min} = 2,\ |Q|_{min} = 5,\ \gamma_{min} = 0.5$$

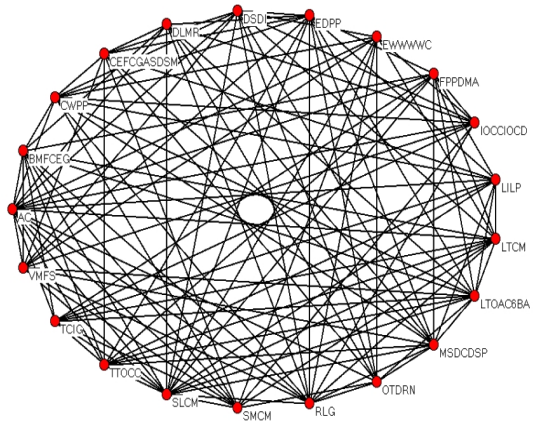
Experimental Results: CiteSeer

top-σ (support)				top-ϵ (structural correlation)				top-δ_{lb} (norm. str. correlation)			
S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}
system paper	57,906	.16	.77	network sensor	3,276	.47	108.7	**node wireless**	**2,086**	**.35**	**164.4**
base paper	56,566	.10	.47	network hoc	2,744	.47	141.2	protocol rout	2,134	.35	157.6
paper result	47,516	.08	.45	ad network hoc	2,725	.44	134.6	memori cach	2,150	.32	143.8
paper model	43,929	.09	.59	network rout	5,084	.41	48.0	network hoc	2,744	.47	141.2
us paper	43,573	.05	.32	network wireless	5,242	.40	44.7	protocol wireless	2,048	.29	138.7

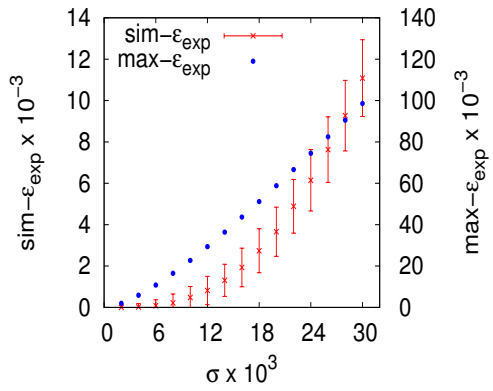
Table: Attribute sets



(a) $\mathcal{G}(\{node, wireless\})$



(b) $\{system, performance\}$



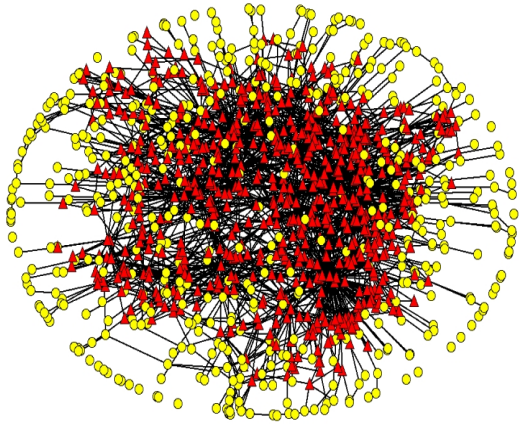
(c) $\sigma \times \epsilon_{exp}$

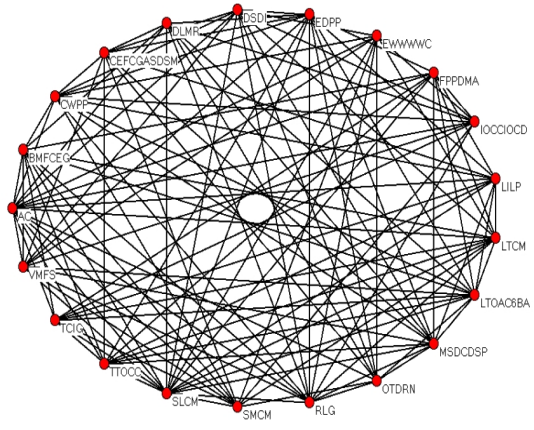
$$\sigma_{min} = 2,000,\ |S|_{min} = 2,\ |Q|_{min} = 5,\ \gamma_{min} = 0.5$$

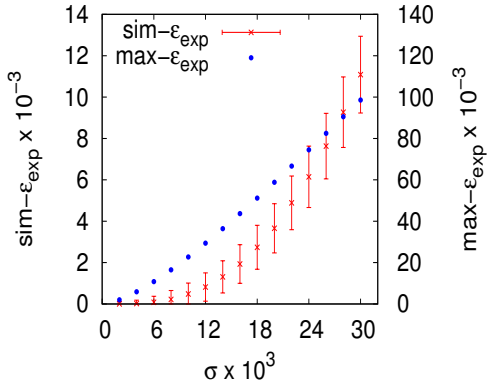
Experimental Results: CiteSeer

top-σ (support)				top-ϵ (structural correlation)				top-δ_{lb} (norm. str. correlation)			
S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}	S	σ	ϵ	δ_{lb}
system paper	57,906	.16	.77	network sensor	3,276	.47	108.7	node wireless	2,086	.35	164.4
base paper	56,566	.10	.47	network hoc	2,744	.47	141.2	protocol rout	2,134	.35	157.6
paper result	47,516	.08	.45	ad network hoc	2,725	.44	134.6	memori cach	2,150	.32	143.8
paper model	43,929	.09	.59	network rout	5,084	.41	48.0	network hoc	2,744	.47	141.2
us paper	43,573	.05	.32	network wireless	5,242	.40	44.7	protocol wireless	2,048	.29	138.7

Table: Attribute sets







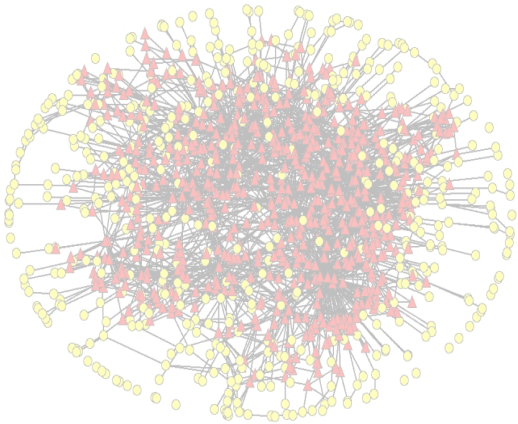
(a) $\mathcal{G}(\{node, wireless\})$ (b) $\{system, performance\}$ (c) $\sigma \times \epsilon_{exp}$

$$\sigma_{min} = 2,000,\ |S|_{min} = 2,\ |Q|_{min} = 5,\ \gamma_{min} = 0.5$$

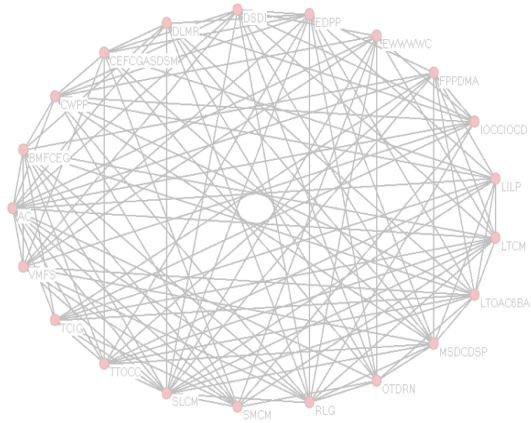
Experimental Results: CiteSeer

Summary:

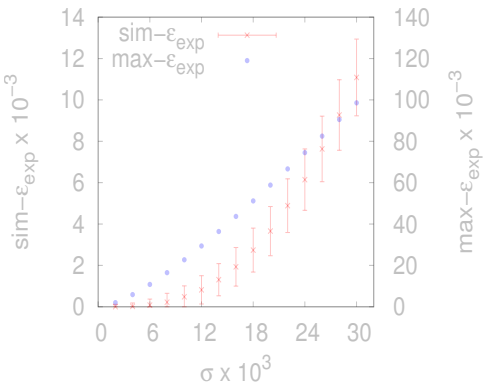
- SCPM = attribute sets X graph structure
- SCPM identifies interesting SCPs (dense subgraphs)
- Our analytical formulation ≈ simulation results
- Statistically significant patterns are more relevant



(a) $\mathcal{G}(\{node, wireless\})$



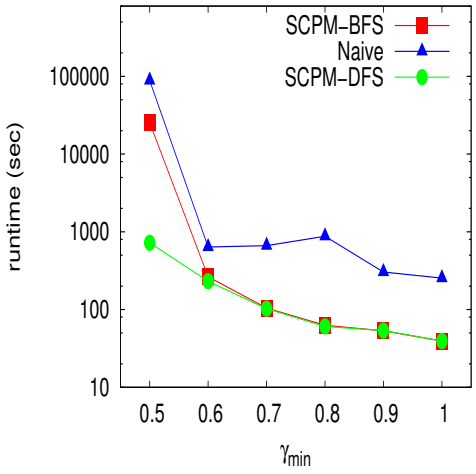
(b) $\{system, performance\}$



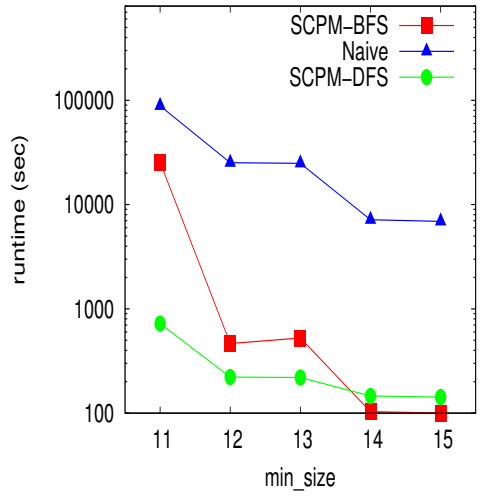
(c) $\sigma \times \epsilon_{exp}$

$\sigma_{min} = 2,000$, $|S|_{min} = 2$, $|Q|_{min} = 5$, $\gamma_{min} = 0.5$

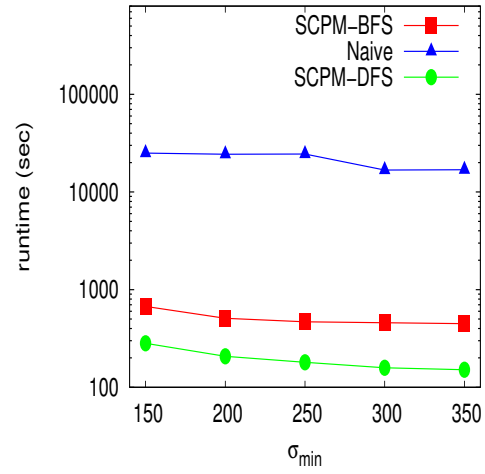
Experimental Results: Performance



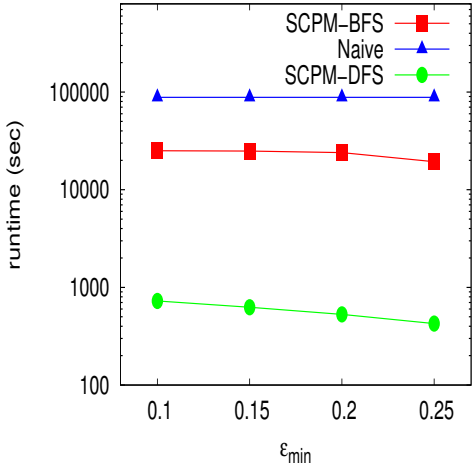
(a) Runtime x γ_{min}



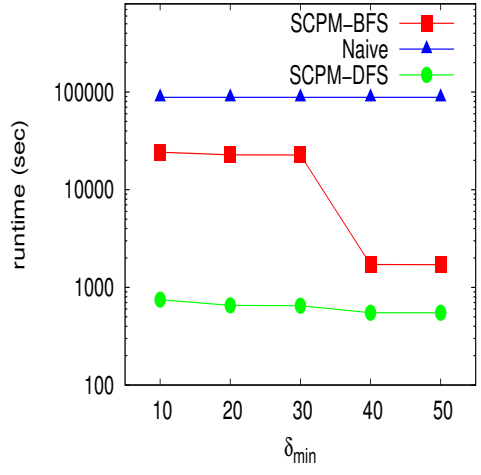
(b) Runtime x *min_size*



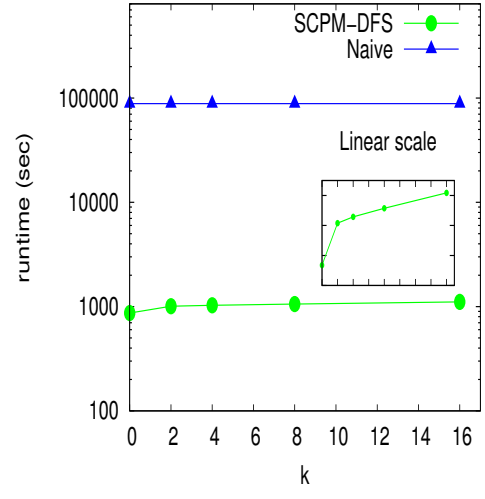
(c) Runtime x σ_{min}



(d) Runtime x ϵ_{min}

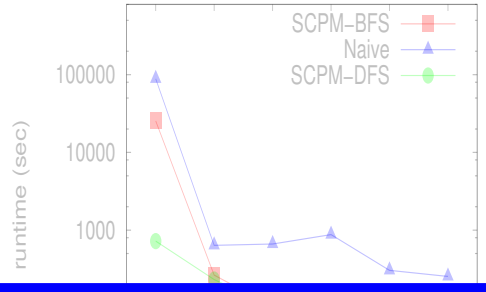


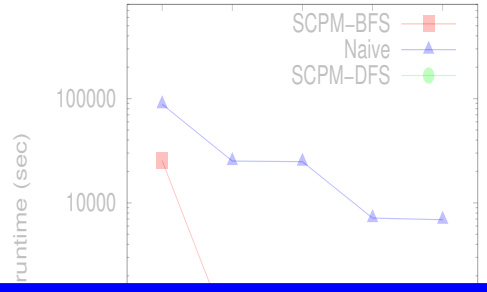
(e) Runtime x δ_{min}

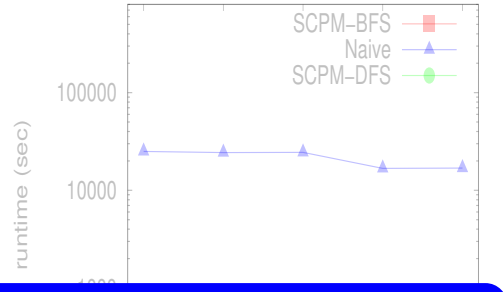


(f) Runtime x k

Experimental Results: Performance





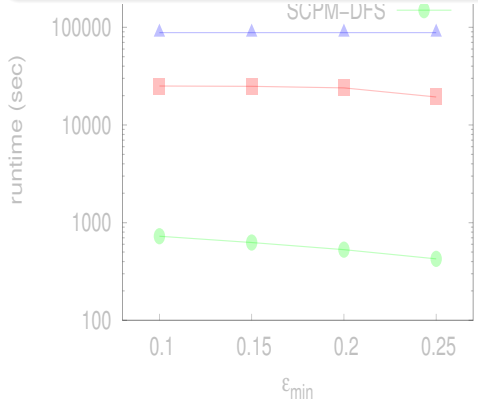


Summary:

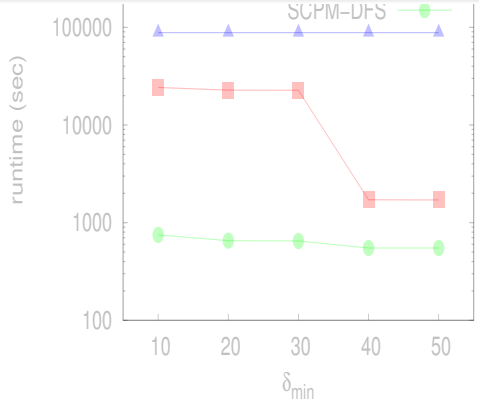
DFS is better than BFS

SCPM outperforms the baseline algorithm

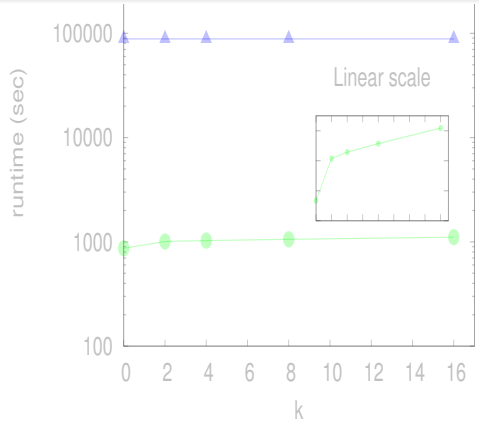
Parameters increase pruning power



(d) Runtime x ϵ_{min}



(e) Runtime x δ_{min}



(f) Runtime x k

Concluding Remarks

Structural correlation pattern mining enables the discovery of relevant correlations between attributes and dense subgraphs

This work has shown that:

- SCPM enables the discovery of:
  - Attribute sets correlated with graph structure
  - Large and dense structural correlation patterns
- Statistiscal significance leads to more relevant patterns
- The vertex and attribute set pruning techniques are effective

Limitations:

- Demands parameter setting
- Performance is still an issue

Future work:

- Applying SCPM to relational learning tasks
- A distributed algorithm for structural correlation mining

Mining Attribute-structure Correlated Patterns in Large Attribute Graphs

More information:

arlei@dcc.ufmg.br
http://www.dcc.ufmg.br/~arlei
http://code.google.com/p/scpm/

This student has been supported by the VLDB student fellowship. Thanks!

Mining Attribute-structure Correlated Patterns in Large Attributed Graphs

@VLDB'12, Istanbul, Turkey

Arlei Silva[1,3], Wagner Meira Jr.[1], Mohammed J. Zaki[2]

[1]Computer Science Department – Universidade Federal de Minas Gerais, Brasil

[2]Computer Science Department – Rensselaer Polytechnic Institute, USA

[3]Computer Science Department – University of California, Santa Barbara, USA