Efficient Structural Clustering on Probabilistic Graphs

Consider an unweighted and undirected probabilistic graph $G = (V, E, P)$ where V is the set of vertices, E is the set of edges, and P denotes the set of probabilities. In G,each edge $e \in E$ is associated with a probability $P_e \in P$

Let $G' = (V, E_G)$ be a possible world which is realized by sampling each edge in G according to the probability P_e clearly we have $E_G \in E$ the probability $Pr[G'|G]$ of sampling this possible world is calculated as

$P_r[G'|G] = \prod_{e \in E_G} P_e \prod_{e \in E \setminus E_G} (1 - P_e)$

we make use of $G' \sqsubseteq G$ to indicate that G is a possible world of G' clearly there are a total of $2^{|E|}$ possible worlds in graph G because each edge provides a binary sampling decision. for convenience,we sue a notation G to denote a probabilistic graph,and utilize a notation G' to denote a possible world or a deterministic graph.

Definition

1. **Structural Neighborhood**

   Given a deterministic graph $G = (V, E_G)$ the structural neighborhood of a vertex $u \in V$ denoted by $N[u]$ is the closed neighborhood of u

   $the\ Structural\ neighborhood\ of\ a\ vertex\ includes\ itself$

2. **Structural Similarity**

   Given a deterministic graph $G = (V, E_G)$ the structural similarity between vertices u and v,denoted by $\sigma(u, v)$ is defined as the number of common vertices in $N[u]$ and $N[v]$ normalized by $|N[u] \cup N[v]|$

   $\sigma(u, v) = \frac{|N[u] \cap N[v]|}{|N[u] \cup N[v]|}$

3. ϵ**-Structural Similarity**

   Given any two neighbor vertices u,v, and a similarity threshold ϵ ,u is structural similar to v in a deterministic graph G if $\sigma(u, v) \geq \epsilon$ and $e = (u, v) \in E_G$

Problem Formulation

Definition

1. *Probability of Structural Similarity*

   Given a similarity threshold $0 < \epsilon \leq 1$ the probability of structural similarity that $\sigma(e) \geq \epsilon$ is defined as the sum of the probabilities of all the possible worlds $G' \sqsubseteq G$ such that the structural similarity of $e = (u, v)$ is no less than ϵ in each possible world G

   $Pr[e, \epsilon] = \sum_{G' \sqsubseteq G} Pr[G'|G] \cdot I(\sigma(e) \geq \epsilon)$

   where $I(\sigma(e) \geq \epsilon)$ is a indicator function which equals 1 if $\sigma(e) \geq \epsilon$ and 0 otherwise. If $e \notin E_G$ $I(\sigma(e) \geq \epsilon) = 0$

2. **Reliable Structural Similarity**

   Given an edge $e = (u, v)$ and a threshold η, u is called reliable structural similar to v if $Pr[e, \epsilon] \geq \eta$

3. (ϵ, η)-**Reliable Neighborhood**

   Given a similarity threshold $0 < \epsilon \leq 1$ and a probability threshold $0 < \eta \leq 1$ the (ϵ, η) -reliable neighborhood of u is defined as the subset of vertices in $N[u]$ such that $Pr[e = (u, v), \epsilon] \geq \eta$

A vertex is termed as a reliable core vertex if it has a sufficient number of reliable similar neighbors.

4. (ϵ, η, μ)**-Reliable Core Vertex**

   Given a similarity threshold $0 < \epsilon \leq 1$,a probability threshold $0 < \eta \leq 1$,and an integer $\mu > 2$ a vertex u is a (ϵ, η, μ)-reliable core vertex if $|N_{(\epsilon,\eta)}[u]| \geq \mu$

5. **Reliable Structural reachable**

   Given parameters $0 < \epsilon \leq 1$,$0 < \eta \leq 1$ and $\mu \geq 2$ vertex v is a reliable structural reachable form vertex u if there is a sequence of vertices $v_1, v_2, \ldots, v_l \in V(l \geq 2)$ such that

   - $v_1 = u$ and $v_l = v$
   - $v_1, v_2, \ldots, v_{l-1}$are reliable core vertices;
   - $v_{i+1} \in N_{\epsilon,\eta}(v_i)$ for $1 \leq i \leq l-1$

The Probabilistic Graph Clustering Problem Given a probabilistic graph $G = (V, E, P)$and parameters $0 < \epsilon \leq 1$,$0 < \eta \leq 1$ and $\mu \geq 2$ the problem of probabilistic graph clustering is to compute the set $\mathbb{C}$ of reliable clusters in G,Each reliable cluster $C \in \mathbb{C}$ should have at least two vertices and satisfy:

- **Maximality** for each reliable core vertex $u \in C$,all vertices that are reliable structure-reachable from u must belong to C;
- **Connectivity** for any two vertices $v_1, v_2 \in C$ there existed a vertex $u \in C$ such that both v_1 and v_2 are reliable structure-reachable from u

6. **Hub and Outlier**

   Given the set of $\mathbb{C}$ of reliable clusters in a probabilistic graph G,a vertex u that not in any reliable cluster in $\mathbb{C}$ is a hub vertex if it connects two or more reliable clusters,and it is an outlier vertex otherwise.

for each $e = (u, v) \in G$ the number of possible values of the structural similarity

between u and v over all the possible worlds can be bounded by $O(k_{join} \times k_{union})$,where $k_{join} = |\overline{N}[u] \cap \overline{N}[v]|$ and $k_{union} = |\overline{N}[u] \cup \overline{N}[v]|$

记$N(u) = N[u] \setminus u$ 记$N(v) = N[v] \setminus v$ 记$NV[e] = N(u) \cup N(v)$ 将$NV[e]$中的顶点按照顶点ID进行排序。

对于一条边$e = (u, v)$,$\sigma(e) = \frac{m}{n}$

现在记$NV'[e] = \{u, v\}$ $X(0, 2, 2) = P_e$,

现在按顺序在$NV[e]$中取第h个元素$w_h = NV[e][h]$，将其加入到$NV'[e]$中，对于新加入的顶点，顶点u,v和该顶点分别存在连接$e_1 = (u, w_h)$ $e_2 = (v, w_h)$

对于下面的三种情况

1. 边e_1 , e_2 同时存在，则在这一步中，$m_{new} = m_{pre} + 1$ $n_{new} = n_{pre} + 1$ 概率 $X(h+1, m_{new}, n_{new}) = P_{e_1} P_{e_2} X(h, m_{pre}, n_{pre})$
2. 边e_1 ,e_2 只有一条边存在，在该步中，$m_{new} = m_{pre}$,$n_{new} = n_{pre} + 1$

   $$X(h+1, m_{new}, n_{new}) = (P_{e_1}(1 - P_{e_2}) + (1 - P_{e_1})P_{e_2})X(h, m_{pre}, n_{pre})$$
3. 边e_1和e_2 都不存在，在该步中，$m_{new} = m_{pre}$, $n_{new} = n_{pre}$

   $$X(h+1, m_{new}, n_{new}) = (1 - P_{e_1})(1 - P_{e_2})X(h, m_{pre}, n_{pre})$$

   综上

   $$\begin{aligned} X(h+1, m_{new}, n_{new}) = & P_{e_1} P_{e_2} X(h, m_{pre}, n_{pre}) \\ & + (P_{e_1}(1 - P_{e_2}) + (1 - P_{e_1})P_{e_2})X(h, m_{pre}, n_{pre}) \\ & + (1 - P_{e_1})(1 - P_{e_2})X(h, m_{pre}, n_{pre}) \end{aligned}$$

伪代码如下：

Input: G=(V,E,P),an edge $e=(u,v)\in E$,and similarity threshold ϵ

Output: the probability $Pr(e,\epsilon)$ when the structural similarity of e is no less than ϵ

1. Initialize $X(h,m,n)\leftarrow 0$,for all $h\in[0,k_{union}]$,$m\in[0,k'_{join}]$,and $n\in[0,k'_{union}]$.
2. $Pr(e,\epsilon)\leftarrow 0$
3. $X(0,2,2)\leftarrow 1$
4. **for** h in $range(1,k_{union}-2)$:
5. $e_1=(u,w_h)$,$e_2=(v,w_h)$
6. **for** n in $range(2,k'_{union}-2)$:
7. **for** m in $range(2,min(n,k'_{join}))$:

$$\begin{aligned}X(h,m,n)=&\,p_{e_1}p_{e_2}X(h-1,m-1,n-1)\\&+((1-p_{e_1})p_{e_2}+p_{e_1}(1-p_{e_2}))X(h-1,m,n-1)\\&+(1-p_{e_1})(1-p_{e_2})X(h-1,m,n)\end{aligned}$$

8. **for** n in $range(2,k_{union})$:
9. **for** m in $range(\lceil n\epsilon\rceil),min(n,k_{join})$:
10. $Pr(e,\epsilon)=Pr(e,\epsilon)+X(k_{union}-2,m,n)$
11. **return** $P_e*Pr(e,\epsilon)$

时间复杂度分析

对于一条边e计算$P(e,\epsilon)$ 的时间复杂度是$O(k^2_{union}k_{join})$ $k_{union}\le 2*d_{max}$

$k_{join}\le min\{d_u,d_v\}$,原式的上界为$O(d^2_{max}\times\sum_{(u,v)\in E}min\{d_u,d_v\})=O(d^2_{max}\times\alpha\times m)$

α denotes the arboricity of the graph G and $m=|E|$

空间复杂度

在计算时，需要使用一个二维矩阵存储值，空间复杂度为$O(k_{union}k_{join})$

Optimization

1. **Basic Pruning Rules**
   - **Pruning Improper Edges**

     For any edge $e=(u,v)\in E$ if $P_e<\eta$ we have $Pr[e,\epsilon]<\eta$
   - **Avoiding Duplicate Computation**

     For any edge $e=(u,v)\in E$,$Pr[e=(u,v),\epsilon]=Pr[e=(v,u),\epsilon]$ always holds
2. **Early Termination**

   在第h次的计算中，如果$N[u]\cap N[v]$ 中的所有节点都被处理，如果此时小于ϵ 则在处理非公共节点的时候，不会再大于ϵ

Algorithm 3. Improved DP for Computing $\Pr(e, \epsilon)$

if $p_e < \eta$ **then**
  return; `/* Property 1 pruning rule */`
`Lines 1-3 in Algorithm 2;`
for $h \leftarrow 1$ **to** $k_{union} - 2$ **do**
  for $m' \leftarrow 2$ **to** $\min\{h+2, k_{join}\}$ **do**
    $\tau \leftarrow \frac{m'}{\epsilon}$;
    for $n' \leftarrow m'$ **to** $\min\{h+2, k_{union}\}$ **do**
      if $h > k_{join} - 2$ *and* $n' \geq \tau$ **then**
        break; `/* early termination */`
      $X(h, m', n') \leftarrow p_{(w_h,u)} p_{(w_h,v)} X(h-1, m'-1, n'-1) + ((1-p_{(w_h,u)})p_{(w_h,v)} + p_{(w_h,u)}(1-p_{(w_h,v)}))X(h-1, m', n'-1) + ((1-p_{(w_h,u)})(1-p_{(w_h,v)}))X(h-1, m', n')$
`Lines 8-11 in Algorithm 2;`

3. **Pruning by Lower and Upper Bounds**

对于分母来说，$n \leq k_{union}$ 如果固定分母为k_{union} 我们能够得到$Pr[e, \epsilon]$ 的下界。

则在变量 $X(h, m, n)$中, n 不起作用，源公式可简化为

$$X(h, m) = P_{e_1} P_{e_2} X(h-1, m-1) + (((1-P_{e_1})P_{e_2} + P_{e_1}(1-P_{e_2})) + (1-P_{e_1})(1-P_{e_2}))X(h-1, m)$$

对于分子来说，$m \leq k_{join}$ 固定分子为k_{join},可以得到$Pr[e, \epsilon]$ 的上界

在变量$X(h, m, n)$的计算中，m不起作用，原公式可以简化为

$$X(h, n) = (P_{e_1} P_{e_2} + P_{e_1}(1-P_{e_2}) + P_{e_2}(1-P_{e_1}))X(h-1, n-1) + (1-P_{e_1})(1-P_{e_2})X(h-1, n)$$

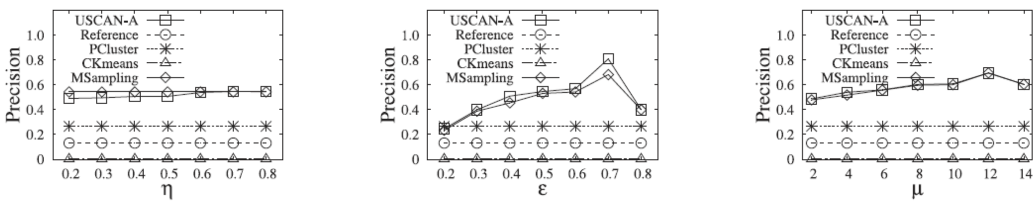
一个约束更强的上界

将$V = (N[u] \cup N[v]) \setminus (N[u] \cap N[v])$ 中的节点去除，不参与$Pr[e, \epsilon]$ 的计算。

因为将V中的节点加入到集合中，会增加分母，减小数值。

实验结果

1. Clustering Precision with Varying Parameters



(a) Vary η (b) Vary ϵ (c) Vary μ

2. Average Expected Density of Different Algorithms.

$AED = \frac{1}{n'} \times \sum_{i=1}^{n'} \sum_{e_j \in E_i} p(e_j) \times /(|V_i| \times (|V_i| - 1))$

n'是聚类的个数

E_i 是第i 个类的边的个数

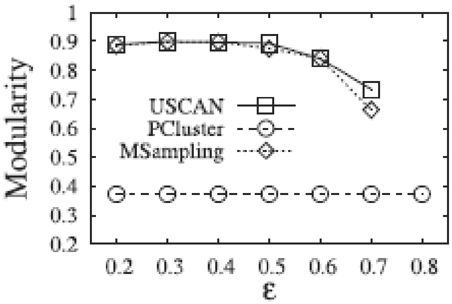
V_i 是第i 个类的点的个数

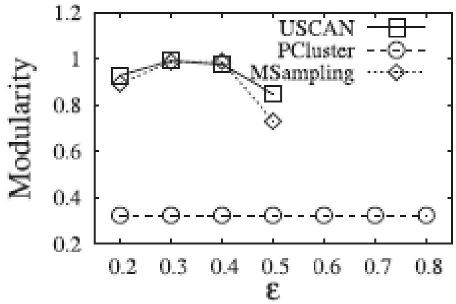
3. Expected Modularity of Various Algorithms

$\overline{Q} = \frac{1}{N} \times \sum_{G' \in G} Q_{G'}$

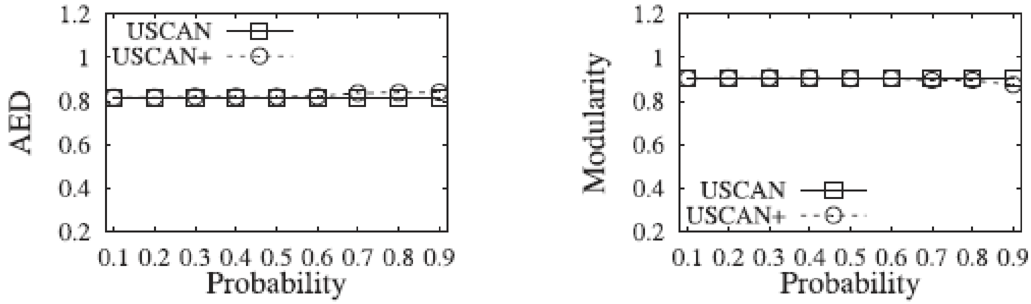
G' 是概率图G 的一种可能，N 是G中可能的个数

$Q_{G'}$ 是modularity of G



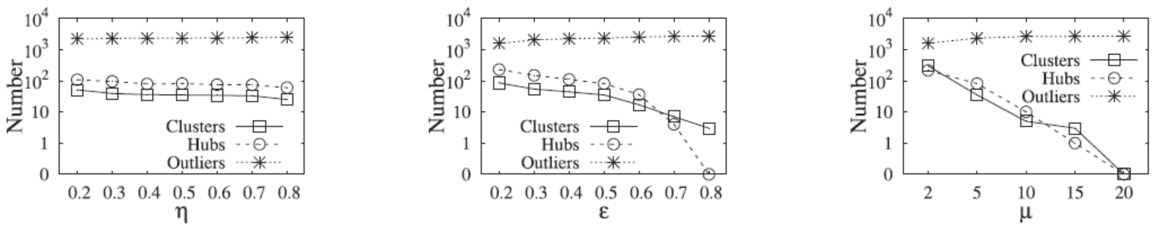


4. Sensitive Analysis

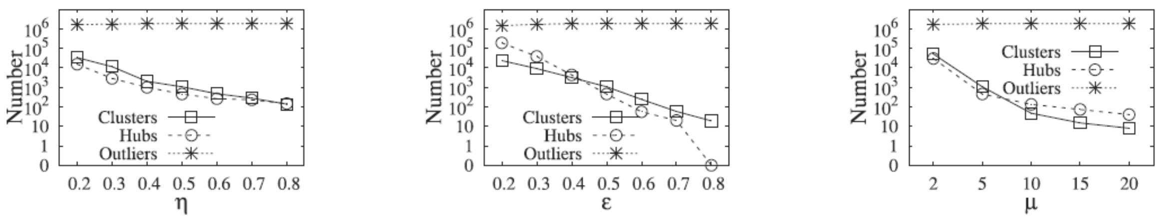


(a) AED (b) Modularity

5. Statistics of the Reliable Structural Clustering.

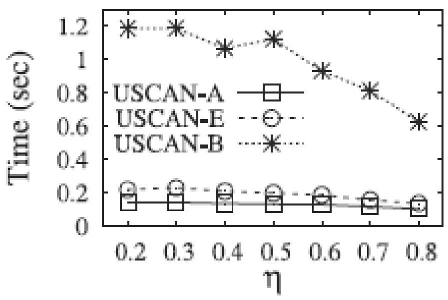


(a) Vary η (CORE) (b) Vary ϵ (CORE) (c) Vary μ (CORE)

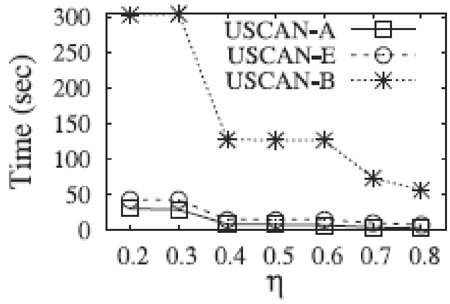


(d) Vary η (DBLPAll) (e) Vary ϵ (DBLPAll) (f) Vary μ (DBLPAll)

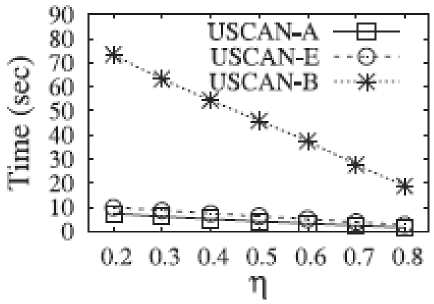
运行时间



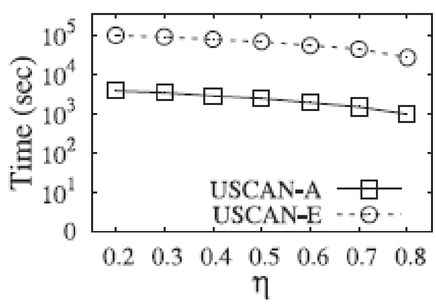
(a) CORE



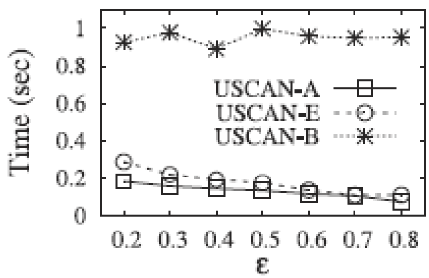
(b) DBLP01



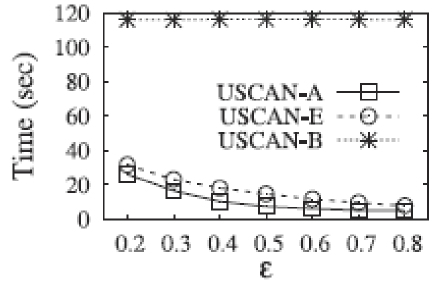
(c) Amazon



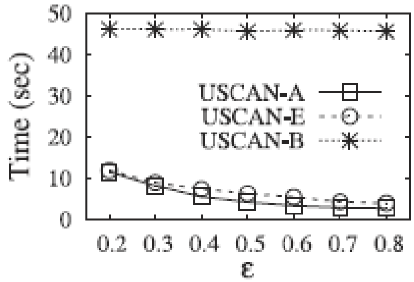
(d) Youtube



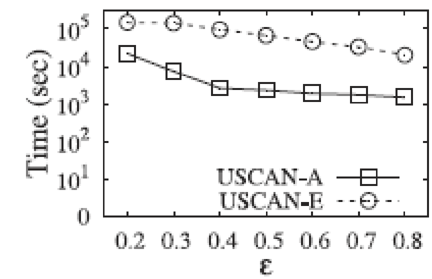
(a) CORE



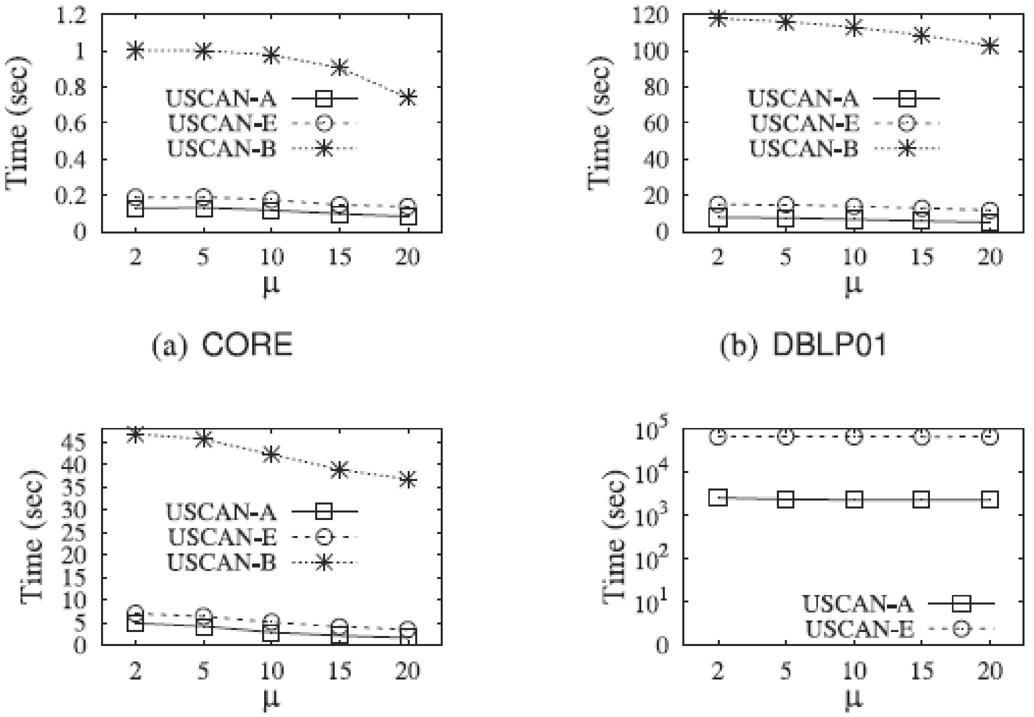
(b) DBLP01



(c) Amazon



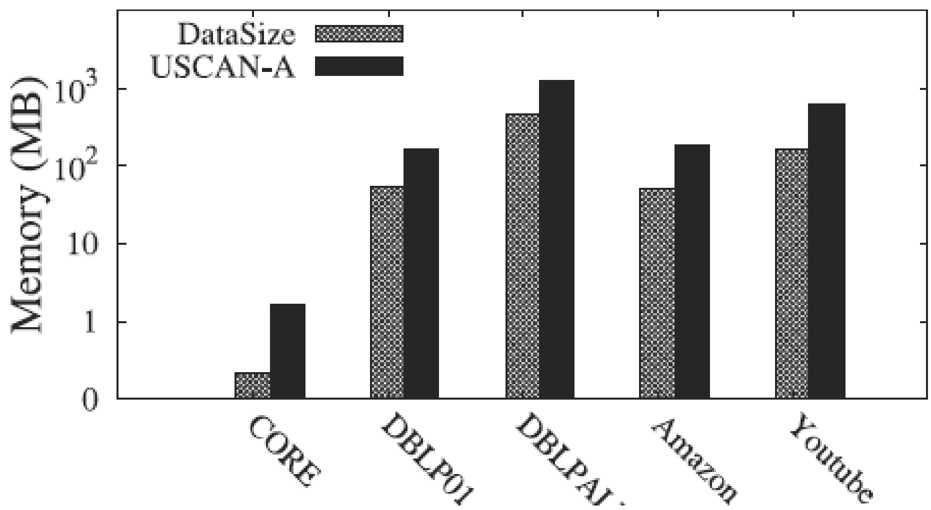
(d) Youtube



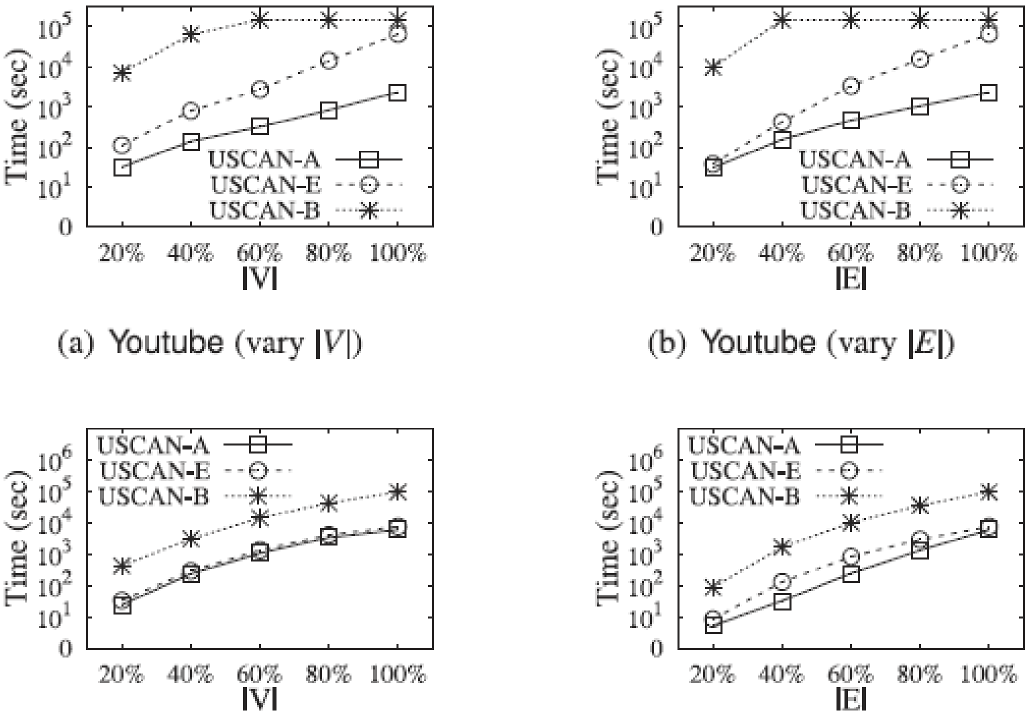
(a) CORE (b) DBLP01

(c) Amazon (d) Youtube

内存消耗



可扩展性



(a) Youtube (vary $|V|$)

(b) Youtube (vary $|E|$)

(c) DBLPAll (vary $|V|$)

(d) DBLPAll (vary $|E|$)