

عنوان ارائه:

یک الگوریتم حاشیه‌سازی افزایشی مبتنی بر افرازبندی برای  
گمنام‌سازی داده‌های جریانی گم شده

**Partitioning based incremental marginalization algorithm for  
anonymizing missing data streams**

توسط: علیرضا صادقی نسب

استاد: دکتر حسین غفاریان

تاریخ ارائه: ۱۳۹۹/۱۲/۲۷

## اطلاعات نویسندگان ■



### Ankhbayar Otgonbayar

[University of the west of Scotland](#)

Verified email at uws.ac.uk

[Internet Of Things](#) [Artificial Intelligence](#) [Deep learning](#) [Data stream processing](#)  
[Privacy preservation](#)

Citations	70	64
h-index	3	3
i10-index	3	3



### Zeeshan PERVEZ

[University of the West of Scotland](#)

Verified email at uws.ac.uk - [Homepage](#)

[Internet-of-Things](#) [Cyber Security](#) [Secure Cloud Services](#) [Data Stream Processing](#)  
[Data Analysis](#)

Citations	939	631
h-index	17	13
i10-index	33	20



### Keshav Dahal

Professor in Intelligent Systems at the [University of the West of Scotland](#)

Verified email at uws.ac.uk

[Intelligent Systems](#) [Artificial Intelligence](#) [Optimisation](#) [Scheduling](#) [Operational Research](#)



Citations	3294	1652
h-index	31	19
i10-index	74	45

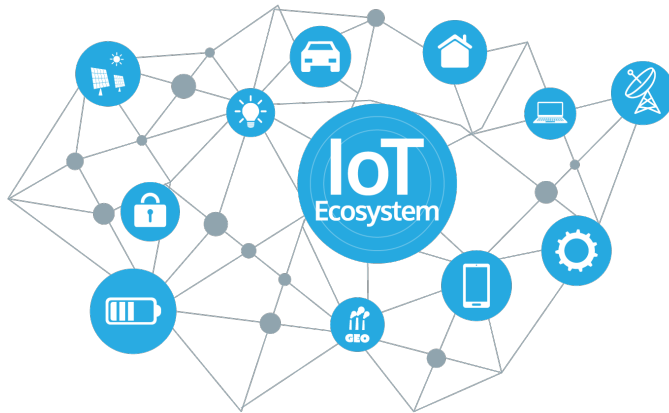
# فهرست مطالب

- مقدمه
- تعاریف پایه
- معرفی روش
- ارزیابی روش
- بررسی نقاط قوت و ضعف

## مقدمه

### ■ اینترنت اشياء

- به مجموعه وسایل مرتبط به هم که روی بستر شبکه بی سیم، داده جمع آوری و ارسال می کنند
- به یک بخش جدایی ناپذیر در جهان مدرن تبدیل شده است
- کاربرد بسیاری در زندگی دارند
- ★ بهینه سازی کنترل ترافیک
- ★ خانه های هوشمند
- ★ کشاورزی و منابع آبی
- ★ حوزه بهداشت و سلامت جامعه




## مقدمه

- لزوم حفظ حریم خصوصی در سامانه‌های اینترنت اشیا
  - از آنجایی که در اغلب موارد، داده‌ها از دستگاه‌های کاربران جمع‌آوری می‌شود، شناسایی و دستیابی به آن‌ها توسط یک شخص مخرب، بسیار زیان‌آور خواهد بود
  - اهمیت داده‌ها می‌تواند از منظر کاربر (شخصی) باشد و یا بیزینسی باشد
  - شخص مخرب می‌تواند به بکارگیری داده‌های دیگر، رفتارهای کاربران را بیاموزد
  - محبوب‌ترین راه‌حل برای پاکسازی داده‌ها، گمنام‌سازی است



## مقدمه

گمنام‌سازی در داده‌ها 

☆ گمنام‌سازی پایگاه‌داده‌ای

✓ فرآیند بر روی یک مجموعه‌داده جمع‌آوری شده ثابت انجام می‌شود

✓ هدف اصلی، کاهش میزان از دست دادن اطلاعات است

☆ گمنام‌سازی داده‌های جریانی

✓ در داده‌های جریانی، زمان بسیار مهم است زیرا داده‌ها پس از یک میزان تاخیر، منقضی

شده و بلااستفاده می‌شوند

✓ فرآیند گمنام‌سازی به صورت پویا انجام می‌شود



Publication delay  Information loss

## مقدمه

■ صفتهای بانکهای اطلاعاتی به ۴ دسته تقسیم می شوند:

- صفت شناسه
- صفت شبه شناسه
- صفت حساس یا محرمانه
- صفت غیر حساس یا غیر محرمانه

Key Attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

## مقدمه

### ■ الگوریتم $K - Anonymity$ :

- اگر یک سابقه در جدول یک مقدار صفت شبه شناسه داشته باشد؛ حداقل  $K - 1$  سابقه دیگر همان مقدار صفت شبه شناسه را خواهند داشت
- به زبان دیگر، حداقل اندازه گروه در صفت شبه شناسه،  $K$  خواهد بود

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV



## مقدمه

★ گم شدن داده در جریان داده اینترنت اشیاء؛ چرا و چطور

★ اینترنت اشیاء شامل دستگاه‌ها بسیار گوناگون و با مشخصات متفاوت از هم هست

★ هیچ الگو استفاده تعریف شده جهانی برای دستگاه‌های اینترنت اشیاء وجود ندارد، بنابراین این

دستگاه‌ها ممکن است داده‌هایی رو تولید کنند که حاوی مقادیر گم شده باشند

★ ۴ دلیل اصلی وجود (تولید) داده‌های گم شده:

☑ تنظیمات متفاوت در دستگاه‌های کاربران

☑ الگوی استفاده متفاوت

☑ شرایط محیطی غیرقابل پیش‌بینی

☑ کنترل اشتراک اطلاعات شخصی



## مقدمه

با داده‌های گم شده چکار کنیم؟

★ نسبت‌دادن یا *imputation*: جهت بازسازی داده‌های گم شده، مقادیر مناسب از پیش محاسبه شده جایگذاری مقدار گم شده می‌شود

★ حاشیه‌سازی یا *marginalization*: در این حالت، به داده اصلی، چیزی اضافه نمی‌شود. در نتیجه از لحاظ تحلیلی، از حالت قبل بهتر است. با داده‌های گم شده در این حالت، مانند مقدار *NULL* رفتار می‌شود

★ افرازبندی یا *partitioning*: در این حالت، مجموعه داده به چند زیرمجموعه بدون مقادیر گم شده تقسیم و تبدیل می‌شود و سپس روش‌های گمنام‌سازی بر روی هر یک از آن‌ها، اعمال می‌شود. اگر نسبت داده‌های گم شده به حجم داده اصلی زیاد باشد، این روش مقرون به صرفه نخواهد بود

# تعاریف پایه

**Definition 1 (Tuple of IoT data)** *Tuple of IoT data is defined as:  $t(id_t, Q_t, ts_t)$ - where  $id_t$  is the identity of an individual,  $Q_t = \{q_1, q_2, \dots, q_m\}$  is a set of QIDs of a tuple, and  $ts_t$  is a time-stamp of the tuple arrival.*

**Definition 2 (Missing data stream)** *Let  $Q$  be the QID's of the data stream, where  $Q = \{q_1, q_2, \dots, q_n\}$ . The missing data streams are defined as:  $S(id, Q_t, ts)$  where  $id$  is the individual's identity,  $Q_t$  is the subset of  $Q$  ( $Q_t \subseteq Q$ ) that describes a receiving tuple, and  $ts$  is the time-stamp of the tuple.*

**Definition 3 (K-Anonymous cluster)** *Let  $C(Q_c)$  be a cluster  $C$  generated out of missing data stream  $S$ . If the  $C(Q_c)$  contains not less than  $K$  number of identities in its composition, then,  $C(Q_c)$  is a  $K$ -anonymous cluster.*

**Definition 4 (Partition on  $Q_p$ )** *Let  $P$  be a set of tuples that shares exact same QIDs set  $P(Q_p) = \{t_1(pid_1, Q_p, ts_1), t_2(pid_2, Q_p, ts_2), \dots, t_z(pid_z, Q_p, ts_z)\}$ , then  $P$  is a partition created on  $Q_p$ .*

# تعاریف پایه

**Definition 5 (Distance between two tuples)** Let  $t_1(pid, Q_1)$  and  $t_2(pid, Q_2)$  be the tuple of missing data stream  $S$ . The distance of  $t_1$  and  $t_2$  is calculated on the common QIDs of both tuples.

$$Distance(t_1, t_2) = \frac{\sum_{q_i \in |Q_1 \cap Q_2|} d_i(q_i)}{|Q_1 \cap Q_2|} \quad (1)$$

$$d_i(q_i) = \begin{cases} \frac{|r_{i.1} - r_{i.2}|}{|R_{i.u} - R_{i.l}|} & \text{if } q_i \text{ is numerical} \\ \frac{|leaves(H_i)| - 1}{|leaves(DGH_i)| - 1} & \text{if } q_i \text{ is categorical} \end{cases} \quad (2)$$

Where  $r_{i.1}(r_{i.2})$  is the value of  $t_1.q_i(t_2.q_i)$  if  $q_i$  is a numeric attribute,  $H_i$  is the lowest common ancestor of  $t_1.q_i(t_2.q_i)$  with respect to  $DGH_i$ .

Clusters generated from missing data streams can contain tuples with different composition of QIDs. Cluster generalization of missing data streams is not similar to traditional cluster generalization. We define the following cluster generalization for such clusters.

# تعاریف پایه

**Definition 6 (Cluster generalization)** Let  $G_j^*(g_1, g_2, \dots, g_m)$  be the generalization of cluster  $C(Q_j)$ . Following calculations are utilized to find the generalization of each QID of  $Q_j$  for generalization.

- 1)  $g_i = [r_{i.min}, r_{i.max}]$ , where  $r_{i.min}(r_{i.max})$  is the min(max) value of  $q_i$  in cluster  $C$ . If  $q_i$  is a numerical.
- 2)  $g_i = H_{i.lowest}$  where  $H_{i.lowest}$  is the lowest common ancestor of the  $q_i$  values of the cluster  $C$ . If  $q_i$  is a categorical.

**Definition 7 (Information loss of tuple)** The information loss of generalizing a tuple  $t(pid, Q_t)$  to  $G_t(g_1, g_2, \dots, g_m)$  is defined as follows:

$$InfoLoss(t, G_t) = \frac{1}{|G_t|} \left( \sum_{q_i \in Q_t} Loss(q_i) \right) \quad (3)$$

Where  $Loss(q_i)$  is the attribute information loss  $q_i$  caused by the attribute generalization  $g_i$ .

$$Loss(q_i) = \begin{cases} \frac{r_{i.u} - r_{i.l}}{R_{i.u} - R_{i.l}} & \text{if } g_i \in [r_{i.l}, r_{i.u}] \\ \frac{|leaves(H_i)| - 1}{|leaves(DGH_i)| - 1} & \text{if } g_i = H \end{cases} \quad (4)$$

Where  $[r_{i.l}, r_{i.u}]$  is the numeric domain of a numeric attribute  $q_i$ , and  $DGH_i$  is the domain generalization hierarchy(DGH) of a categorical attribute  $q_i$ ,  $|leaves(H_i)|$  and  $|leaves(DGH_i)|$  represents the size of a tree rooted on  $H_i$  and  $DGH_i$  respectively.

**Definition 8 (Average information loss)** The average information loss for anonymization of the first  $N$  tuples of missing data stream defined as follows:

$$AverageInfoLoss(N) = \frac{1}{N} \sum_{i=1}^N InfoLoss(t_i, G_i) \quad (5)$$

# معرفی روش

■ نمای کلی الگوریتم

★ متغیر  $S$ ، داده جریانی حاوی مقادیر گم شده است

★ متغیر  $K$ ، درجه الگوریتم  $k$ -گمنامی است

★ متغیر  $\delta$ ، اندازه پنجره کشویی است

★ متغیر  $\omega$ ، اندازه محدودیت زمانی استفاده مجدد

خوشه‌های گمنام شده است

\* بعد از تخصیص پارتیشن به داده  $t$ ، اگر چندتایی فوق در

حال منقضی شدن باشد، اگر اندازه بافر همراه آن بیش از

$K$  باشد، گمنام می‌شود و گرنه از بین خواهد رفت

---

## Algorithm 1 *IncrementalPBM*( $S, K, \delta, \omega$ )

---

```
1: Let be  $Set_p$  a set of partition which will act as buffer,
   initialized empty.
2: Let be  $Set_{kc}$  a set of  $K$ -anonymous cluster, initialized
   empty.
3: while  $S \neq NULL$  do
4:   Read tuple  $t_i$  from  $S$  and assign partition of  $Set_p$  or
   create new partition on it.
5:   if  $t$  is expiring according to  $\delta$  then
6:     Remove expired  $K$ -anonymous clusters' of  $Set_{kc}$ 
   according to  $\omega$ 
7:     Let  $t$  be the expiring tuple
8:     if  $(\sum_{P_i \in Set_p} |P_i|) \geq K$  then
9:       AnonymizationPBM( $t$ )
10:    else
11:      SuppressAnonymization( $t$ )
12:    end if
13:  end if
14: end while
```

---

# معرفی روش

## Algorithm 2 AnonymizationPBM( $t$ )

```
1: Find  $K$ -anonymous cluster  $C_{min}$  from  $Set_{kc}$  which covers  $t$  with minimum information loss
2: if  $C_{min}$  is found then
3:   Use cluster generalization of  $C_{min}$  to publish  $t$ 
4:   RETURN
5: end if
6: Let  $P_t$  be the host partition of the expiring tuple  $t$ 
7: Create temporary partition's set  $S_{pub}$ 
8: Let  $QID_p$  be publication  $QID$  initialized as  $P_t.qid$ 
9: for each  $P_i \in Set_p$  do
10:   if  $P_i.qid \subseteq QID_p$  then
11:     Add  $P_i$  to  $S_{pub}$ 
12:   end if
13: end for
14: do
15:   Find the biggest partition  $P_{sim}$  that is the most similar to  $QID_p$ 
16:    $QID_p = QID_p \cup P_{sim}.qid$ 
17:   for each  $P_i \in Set_p$  do
18:     if  $P_i.qid \subseteq QID_p$  &  $P_i \notin S_{pub}$  then
19:       Add  $P_i$  to  $S_{pub}$ 
20:     end if
21:   end for
22: while  $(\sum_{P_i \in S_{pub}} |P_i|) \geq K$ 
23: Find  $K - 1$  nearest tuple of  $t$  from  $S_{pub}$  and form new cluster  $C_{new}$ 
24: Publish  $C_{new}$  and remove tuples from partitions accordingly
25: Add  $C_{new}$  to  $Set_{kc}$ 
```

■ تابع گمنام‌سازی

★ ابتدا سعی می‌شود خوشه‌ای که کمترین اتلاف اطلاعات را

دارد و شامل  $t$ ، می‌باشد را پیدا کند

★ اگر این خوشه پیدا شود، داده  $t$  و بقیه اعضای خوشه وارد

فرآیند گمنام‌سازی خواهند شد

★ اگر خوشه پیدا نشود، یک پارتیشن ساخته خواهد شد و

داده‌هایی که  $QID$  آن‌ها زیرمجموعه یا مساوی  $QID$  داده

$t$  است، افزوده خواهد شد. در نهایت با استفاده از الگوریتم

$KNN$ ، نزدیک‌ترین داده‌ها انتخاب و در یک خوشه جدید


گنجانده، گمنام و منتشر می‌شوند


# ارزیابی روش

مشخصات ارزیابی 

Algorithm	Parameters
<i>K-VARP</i>	$K=50, \delta=2000, \omega=2000, R=0.2$
<i>IncrementalPBM</i>	$K=50, \delta=2000, \omega=2000$

با الگوریتم مشابه *K-VARP* مقایسه شده است 

از مجموعه داده *Adult* برای ارزیابی عملکرد استفاده شده است 

مشخصات مجموعه داده فوق به صورت زیر است: 

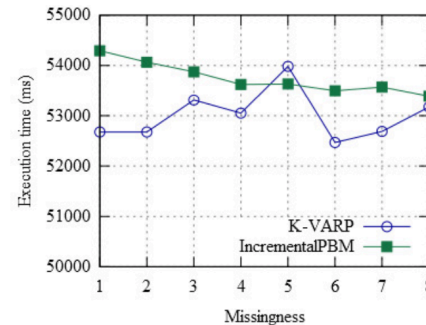
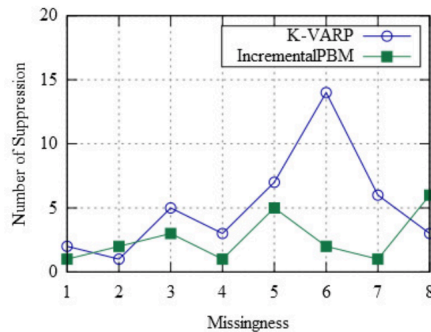
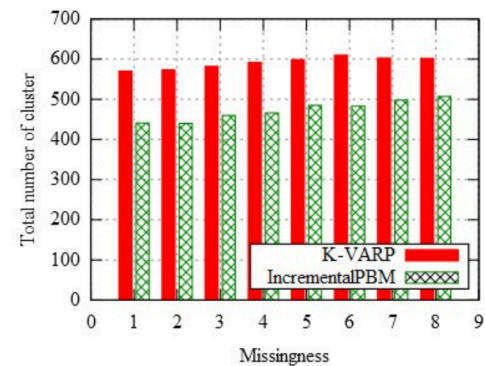
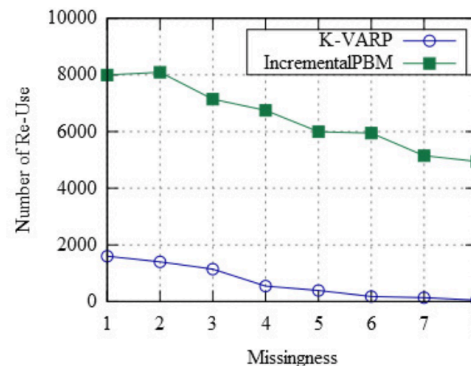
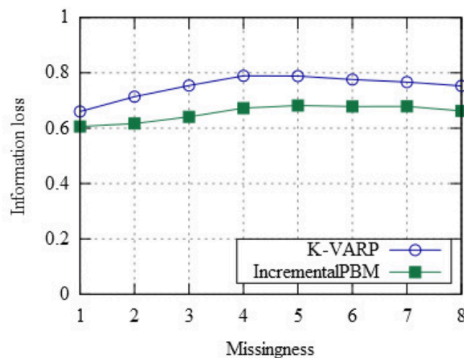
Attribute name	Type	Range	
		Min	Max
Age	<i>Numeric</i>	17	90
Final-weight	<i>Numeric</i>	13769	1484705
Education-number	<i>Numeric</i>	1	16
Capital-gain	<i>Numeric</i>	0	99999
Capital-loss	<i>Numeric</i>	0	4356
Hours-per-week	<i>Numeric</i>	1	99
		Hierarchy tree	
		Height	Nodes
Education	<i>Categorical</i>	5	26
Marital-status	<i>Categorical</i>	4	11
Work-class	<i>Categorical</i>	5	13
Country	<i>Categorical</i>	4	62
Occupation	<i>Categorical</i>	3	15
Relationship	<i>Categorical</i>	3	7
Age	<i>Categorical</i>	3	6
Gender	<i>Categorical</i>	2	3



# ارزیابی روش

نتایج ارزیابی

✓ ارزیابی در مولفه‌های اتلاف اطلاعات، میزان استفاده مجدد خوشه‌ها، تعداد خوشه‌ها، میزان داده‌های حذف شده و زمان اجرا انجام شده است



## بررسی نقاط قوت و ضعف

### ☆ نقاط قوت

☑ مناسب برای سیستم‌های واقعی اینترنت اشیا (حاوی مقادیر گم شده)

☑ ایجاد تعادل مناسب در مولفه‌های اتلاف اطلاعات و میزان زمان اجرا

### ☆ نقاط بهبود

☑ استفاده از روش‌ها و تکنیک‌های هوش مصنوعی در گمنام‌سازی جهت پیشگویی

داده‌های گم شده و تعریف معیار فاصله‌های وفق‌پذیر

## با تشکر از توجه شما