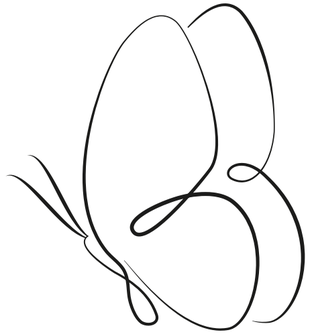
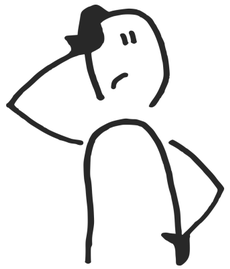


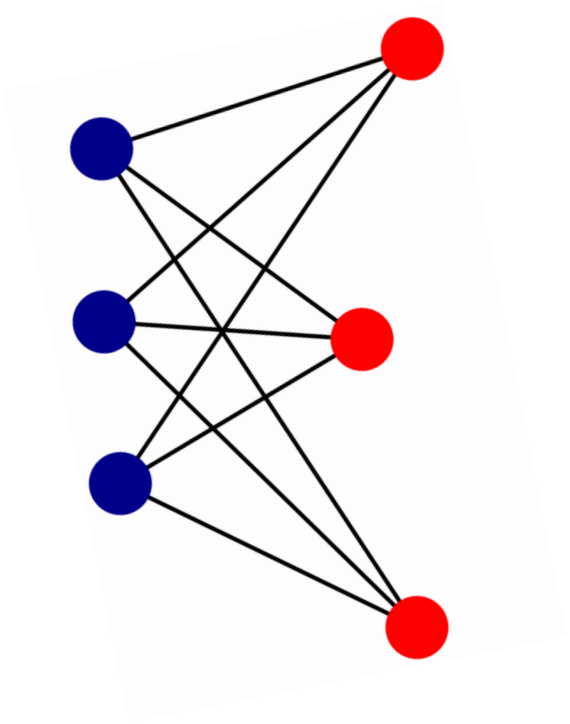
BUTTERFLY COUNTING ON UNCERTAIN BIPARTITE GRAPHS

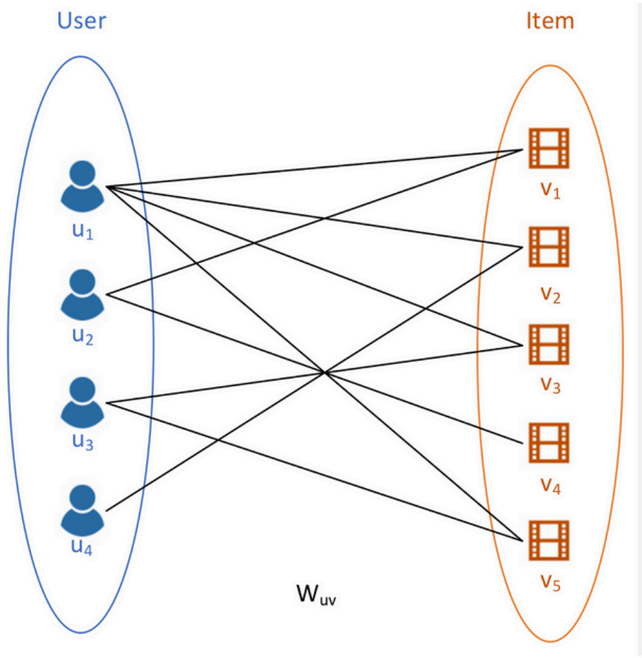
Authors: Alexander Zhou, Yue Wang, Lei Chen
Presented by: Rasheeq Ishmam (RI24C)
Date: April 22, 2025



WHAT IS A BIPARTITE GRAPHS?

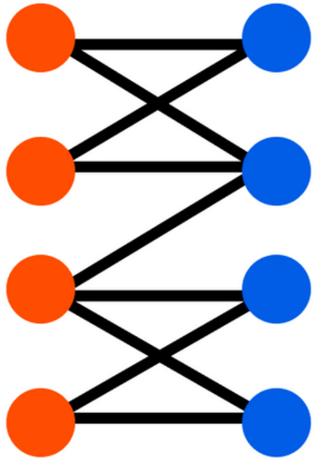
- Bipartite Graph consists of two sets of **nodes**: **Set L** and **Set R**
- **Edges** only connect nodes between these two sets. No edges within the same set.

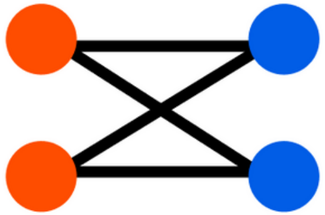


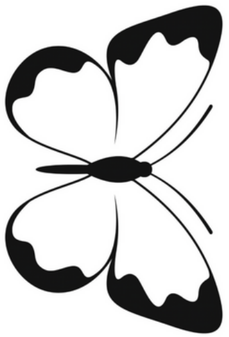


WHAT IS A BUTTERFLY

- A **butterfly** is a subgraph of the bipartite graph.
- A **butterfly** consist of **4 node** with **4 edges** structure in a bipartite graph
- It's a **2x2 biclique**, meaning 2 nodes from Set L are connected to 2 nodes from Set R.
- Represent **clustering/cohesion** between nodes in the graph.

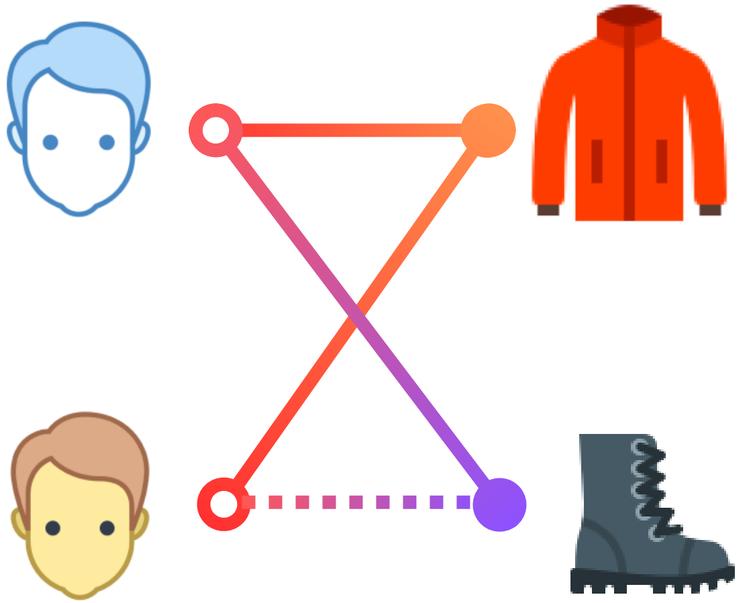






HOW DOES BUTTERFLY WORKS

- **Person 1** and **Person 2** purchased same **Jacket**
- **Person 1** also purchased a pair of **shoe**
- The that pair of shoe can be advertise to person 2



PROBLEM STATEMENT

- **Objective:** Efficiently count uncertain butterflies in a bipartite graph with uncertain edges (edges with probabilities of existence).
- **Challenges:**
  - How to count butterfly subgraphs in graphs with uncertain edges.
  - How can we take only edges that meet a certain **threshold t** are considered for butterfly counting. The **threshold t** allows us to focus on more reliable edges and ignore those with low probability.
  - How can we make it memory efficient.

BASELINE SOLUTION

- The baseline Solution is **Uncertain Butterfly Counting (UBFC)**
- UBFC focuses on removing edges in the bipartite graph that have uncertain existence
  - Edges with **Probability ≥ Threshold** are considered existing and valid for butterfly counting.
  - Edges with **Probability < Threshold** are pruned (removed).
- **Wedge** counting are done in a brute-force manner
- Combining these wedges to form butterflies count.

Algorithm 1: *UBFC*

Input : G: Input Uncertain Bipartite Network
t: Uncertainty Threshold
Output: C_t: Uncertain Butterfly Count

```
1  W_1 ← Extract Backbone Graph;
2  Sort N(u) of each u ∈ V_{W_1} by vertex priority;
3  C_t ← 0;
4  foreach u ∈ V_{W_1} do
5      Create H(w) for each Node w in same partition as
6      foreach v ∈ N(u) : p(v) < p(u) do
7          foreach w ∈ N(v) : p(w) < p(u) do
8              H(w).append(v);
9      foreach Node w : |H(w)| > 1 do
10         foreach Nodes v_1, v_2 ∈ H(w), v_1 ≠ v_2 do
11             if Pr(B(u, w, v_1, v_2)) > t then
12                 C_t ← C_t + 1;
```

PROPOSED SOLUTION

- The proposed Solution is **Improved Uncertain Butterfly Counting (IUBFC)**
- Counts uncertain butterflies by processing pruned **edges** and **wedges**.
- **Steps:**
  - **Edge Pruning:** Remove edges with a probability less than **threshold t**
  - **Wedge Counting:** For each node, count the wedges that can form butterflies using binary search.
  - **Vertex Priority:** Prioritize nodes based on their importance to optimize butterfly counting.
  - **Butterfly Counting:** Combine valid wedges to count uncertain butterflies in the graph.

Algorithm 3: *IUBFC*

Input : G: Input Uncertain Bipartite Network
t: Uncertainty Threshold
Output: C_t: Uncertain Butterfly Count

1 $W_1 \leftarrow$ Extract Backbone Graph;
2 $RemoveUnusableEdges(W_1, t)$;
3 Sort $N(u)$ of each $u \in V_{W_1}$ by vertex priority;
4 $C_t \leftarrow 0$;
5 **foreach** $u \in V_{W_1}$ **do**
6 Create $H(w)$ for each Node w in same partition as u;
7 **foreach** $v \in N(u) : p(v) < p(u)$ **do**
8 **foreach** $w \in N(v) : p(w) < p(u)$ **do**
9 **if** $Pr(\angle(u, v, w)) \geq t$ **then**
10 $H(w).sortedInsert(Pr(\angle(u, v, w)))$;
11 **foreach** $w : |H(w)| > 1$ **do**
12 $C_t \leftarrow C_t + ImprovedListCount(H(w), t)$;

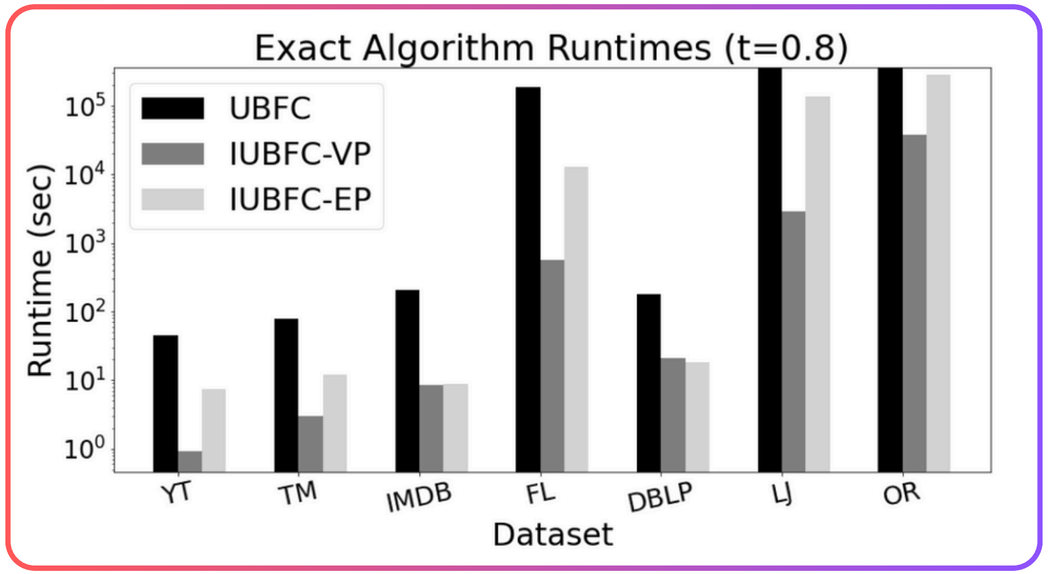
DATASET

- In this research, several real-world datasets is used to evaluate the IUBFC algorithm
- **Dataset:**

Dataset	Edge Prob.	$\|L\|$	$\|R\|$	$\|E\|$	AvgDeg
YouTube (YT)	Uniform	94,238	30,087	293,360	4.719
Teams (TM)	Normal$\tilde{}$(0.8, 0.2)	901,166	34,461	1,366,466	2.921
IMDB	Normal$\tilde{}$(0.6, 0.3)	303,617	896,302	3,782,463	6.229
Flickr (FL)	Uniform	395,979	103,631	8,545,307	34.208
DBLP	Normal$\tilde{}$(0.7, 0.1)	1,953,085	5,624,219	12,282,059	3.241
LiveJournal (LJ)	Normal$\tilde{}$(0.5, 0.2)	3,201,203	7,489,073	112,307,385	21.011
Orkut (OR)	Normal$\tilde{}$(0.5, 0.25)	2,783,196	8,730,857	327,037,487	56.807
CiaoDVD (CD)	Collaborative Filtering	21,019	71,633	*	*
BookCrossing (BC)	Collaborative Filtering	77,802	185,955	*	*

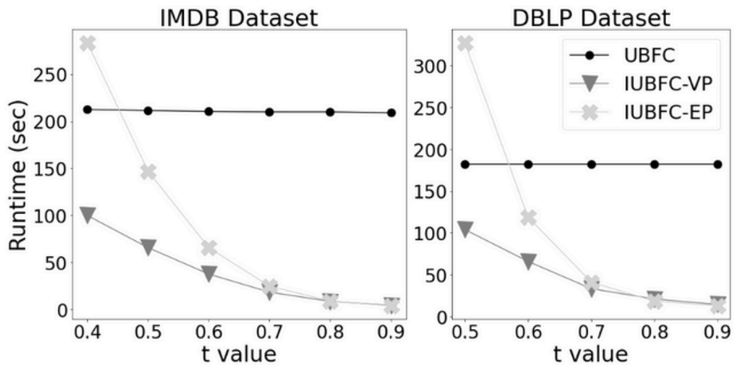
RESULT

- **Evaluation Metrics:**
  - **Runtime:** The total time taken by the algorithm to compute the uncertain butterfly count for each dataset.
  - **Uncertain Butterfly Count:** The number of uncertain butterflies identified by the algorithm
  - **Margin of Error:** Sampling-based algorithms like UBS and PES were used to to quantify the accuracy
  - **Memory Usage:** The amount of memory consumed by the algorithm during computation

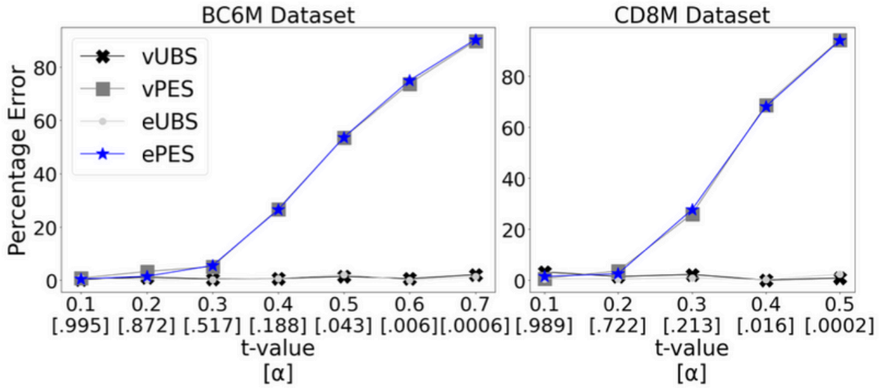


RESULT

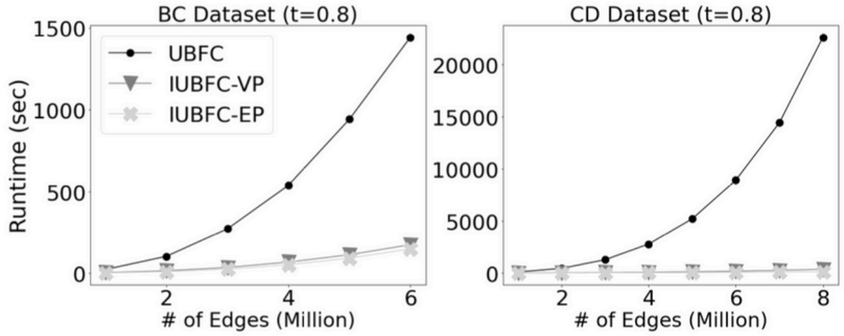
- **Runtime vs Threshold t :**
  - threshold t increases, the runtime decreases



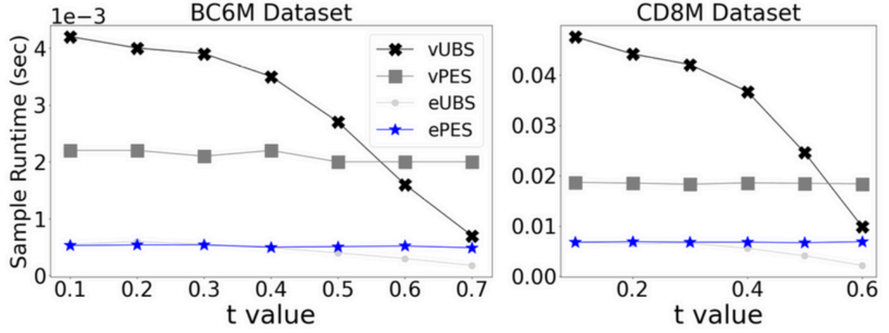
- **Percentage of Error :**
  - The error increases as the threshold t increases



- **Number of edges increase :**
  - Improved algorithm is more stable compared to UBFC



- **Runtime per sample as the threshold value t changes:**
  - ePES is more stability in terms of both runtime and accuracy



IMPLEMENTATION

- **Implementation Result on IMDB Dataset:**

UBFC:

```
IMDB Dataset:

Total Nodes: 1199919
Edges: 3782463

t : 0.4||Number of Uncertain Butterflies: 348028||Runtime: 1816567 ms
t : 0.6||Number of Uncertain Butterflies: 20051||Runtime: 309554 ms
t : 0.7||Number of Uncertain Butterflies: 5295||Runtime: 311195 ms
t : 0.7||Number of Uncertain Butterflies: 5295||Runtime: 311195 ms
```

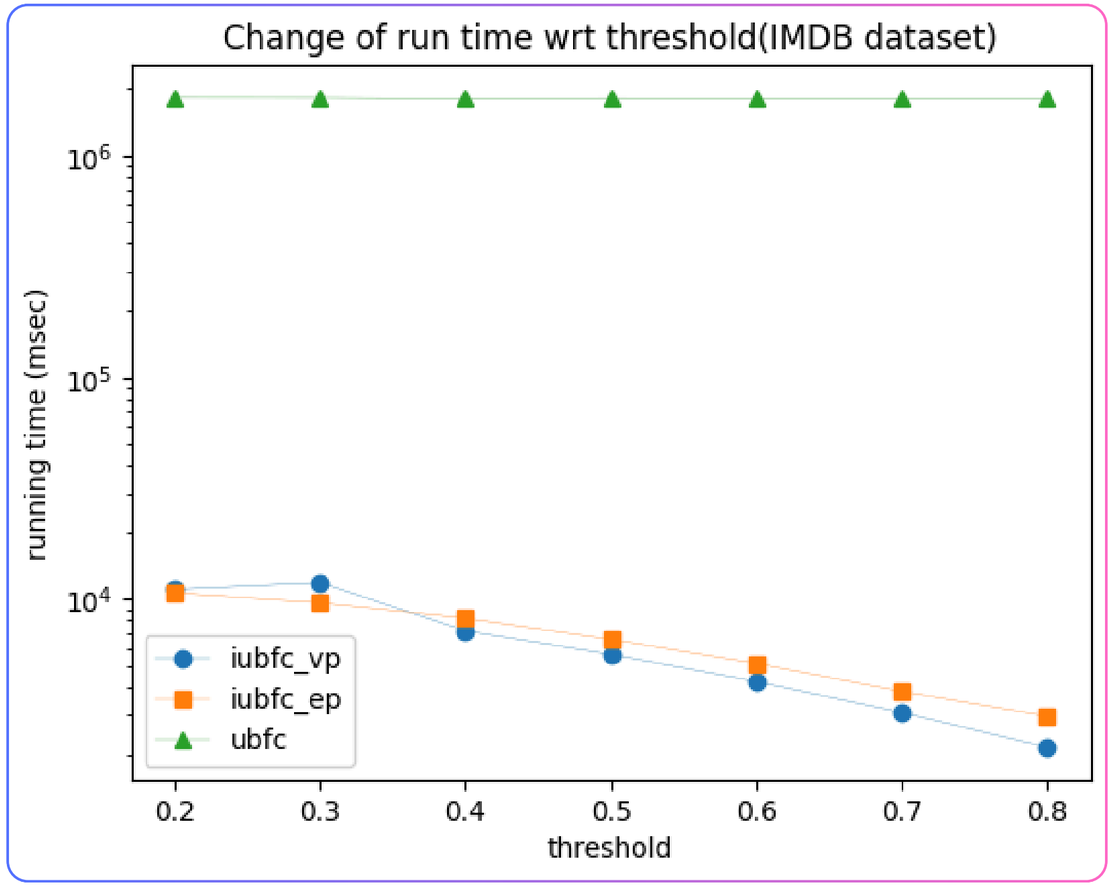
IUBFC:

```
IMDB Dataset:

Total Nodes: 1199919
Edges: 3782463
The number of edges satisfying the threshold: 3497950
t : 0.2||Number of Uncertain Butterflies: 831655||Runtime: 11036 ms
Edges: 3782463
The number of edges satisfying the threshold: 3209475
t : 0.3||Number of Uncertain Butterflies: 215315||Runtime: 11851 ms
Edges: 3782463
The number of edges satisfying the threshold: 2808827
t : 0.4||Number of Uncertain Butterflies: 59888||Runtime: 7185 ms
Edges: 3782463
The number of edges satisfying the threshold: 2308787
t : 0.5||Number of Uncertain Butterflies: 17175||Runtime: 5627 ms
Edges: 3782463
The number of edges satisfying the threshold: 1752095
t : 0.6||Number of Uncertain Butterflies: 4928||Runtime: 4247 ms
Edges: 3782463
The number of edges satisfying the threshold: 1193441
t : 0.7||Number of Uncertain Butterflies: 1205||Runtime: 3079 ms
Edges: 3782463
The number of edges satisfying the threshold: 692348
t : 0.8||Number of Uncertain Butterflies: 235||Runtime: 2150 ms
```

IMPLEMENTATION

- **Runtime vs Threshold plot on IMDB Dataset:**



THANK YOU!