Sampling Representative Users from Large Social Networks

Jie Tang, Chenhui Zhang

Tsinghua University

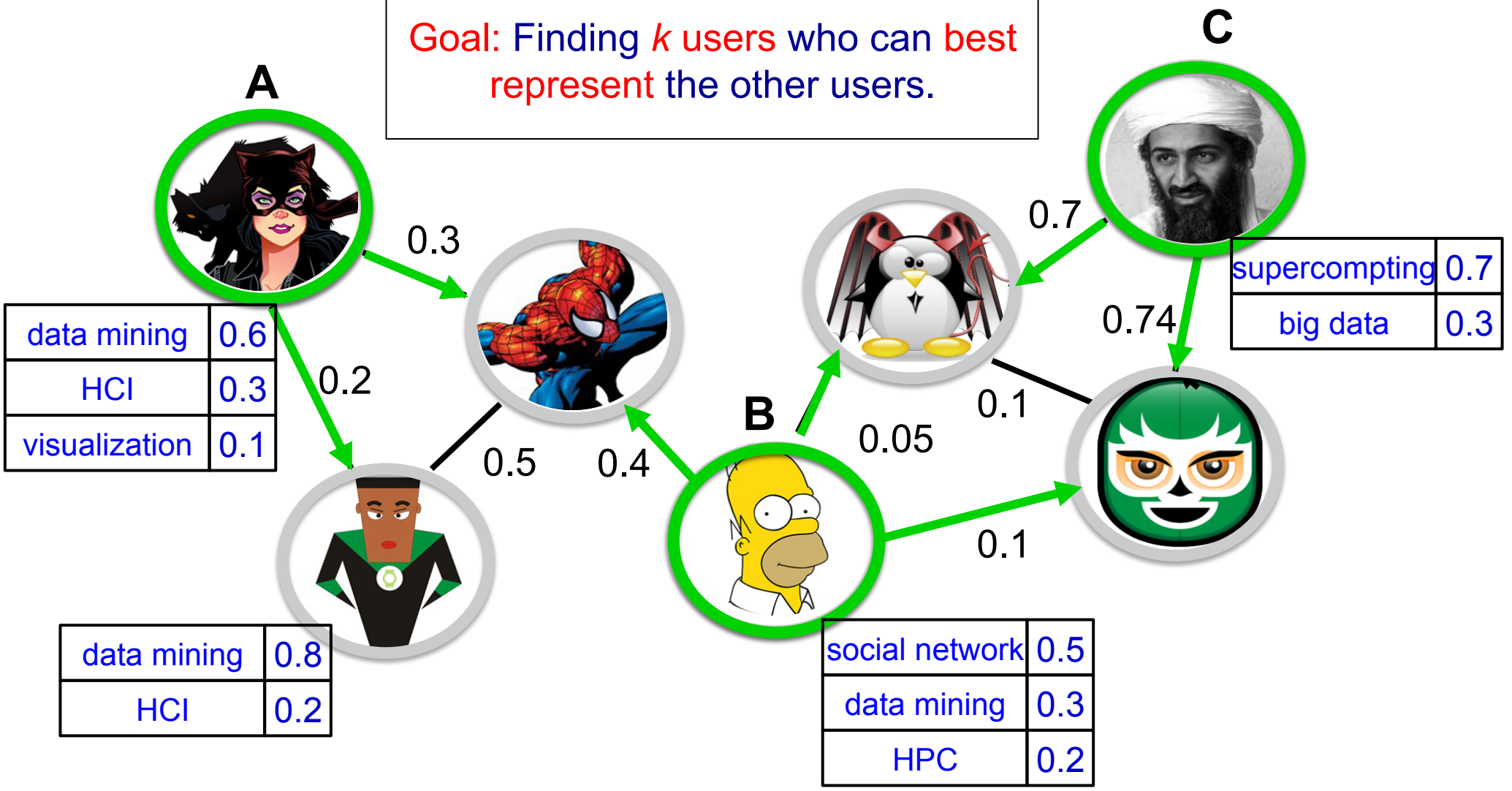
Keke Cai, Li Zhang, Zhong Su

IBM, China Research Lab

Download Code&Data here: http://aminer.org/repuser

Sampling Representative Users

Goal: Finding *k* users who can best represent the other users.

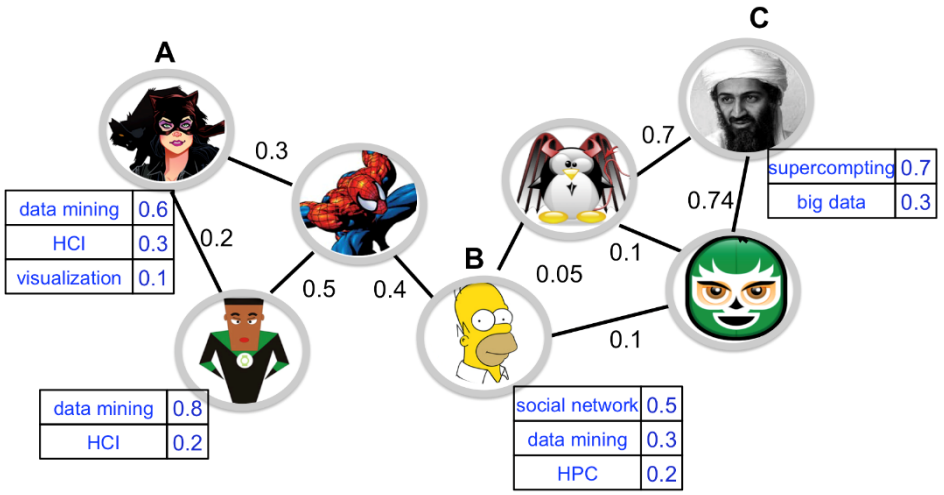


Sampling Representative Users

- Problem Formulation

$$\arg\max_{T \subseteq V, |T|=k} Q(G, X, \mathbf{G}, T)$$

$G=(V, E)$: the input social network



$X=[x_{ik}]_{nxm}$: attribute matrix

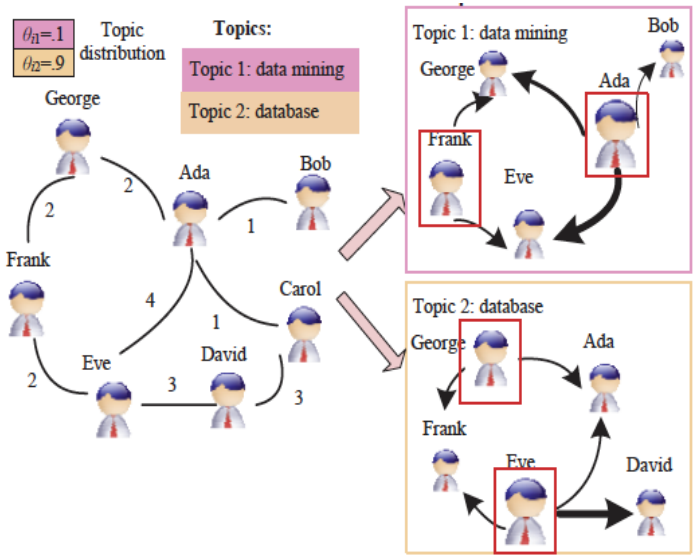
$\mathbf{G}=\{(V_1, a_{j1}), \ldots, (V_t, a_{jt})\}_{jt}$
(V_1, a_{j1}): a subset of users V1 with the attribute value a_{j1}

T ($|T|=k$): Selected subset of representative users

Goal: how to find the subset T ($|T|=k$) in order to maximize the utility function Q?

Related Work

- Graph sampling
  - Sampling from large graphs
    [Leskovec-Faloutsos 2006]
  - Graph cluster randomization
    [Ugander-Karrer-Backstrom-Kleinberg 2013]
  - Sampling community structure
    [Maiya-Berger 2010]
  - Network sampling with bias
    [Maiya-Berger 2011]



- Social influence test, quantification, and diffusion
  - Influence and correlation [Anagnostopoulos-et-al 2008]
  - istinguish influence and homophily [Aral-et-al 2009, La Fond-Nevill 2010]
  - Topic-based influence measure [Tang-Sun-Wang-Yang 2009, Liu-et-al 2012]
  - Learning influence probability [Goyal-Bonchi-Lakshmanan 2010]
  - Linear threshold and cascaded model [Kempe-Kleinberg-Tardos 2003]
  - Efficient algorithm [Chen-Wang-Yang 2009]

The Proposed Models: S3 and SSD

Proposed Models

- **Theorem 1.** *For any fixed Q, the Q-evaluated Representative Users Sampling problem is NP-hard, even there is only one attribute*

$$\arg\max_{T \subseteq V, |T|=k} Q(G, X, \mathbf{G}, T)$$

Prove by connecting to the **Dominating Set Problem**

- Two instantiation models
  - Basic principles: synecdoche and metonymy
  - Statistical Stratified Sampling (S3): Treat one user as being a synecdochic representative of all users
  - Strategic Sampling for Diversity (SSD): Treat one measurement on that user as being a metonymic indicator of all of the relevant attributes of that user and all users

Statistical Stratified Sampling (S3)

- Maximize the representative degree of all attribute groups

$\mathbf{G}=\{(V_1, a_{j1}), \ldots, (V_t, a_{jt})\}_{jt}$
(V_1, a_{j1}): a subset of users V_1 with the attribute value a_{j1}

  - Trade-offs: choose some less representative users on some attributes in order to increase the global representative

- Quality Function

$$Q(G, X, \mathbf{G}, T) = \sum_{(V_l, a_{j_l}) \in \mathbf{G}} m_l \{P(T, l)\}^{\beta}$$

The degree that T represent the l^{-th} attribute group G_l

Avoid bias from large groups

Importance of the attribute group

Approximate Algorithm

Input: The set of users V; the set of attributes A; the attributes values X; the values $R(T, v_i, a_j)$ for every $|T| = 1$ and $v_i \in V$, $a_j \in A$; the number of users to find k.
Output: A set T of k representative users.
$T = \emptyset$;
while $|T| < k$ **do**
  $max = -1$;
  foreach $v_i \notin T$ **do**
    foreach $(V_l, a_{j_l}) \in \mathbf{G}$ *that* v_i *can contribute* **do**
      foreach $v_{i'} \in V_l$ *that* $R(\{v_i\}, v_{i'}, a_{j_l}) > 0$ **do**
        Compute the increment of $R(T, v_{i'}, a_{j_l})$;
      end
      Compute the total increment of $P(T, l)$;
    end
    Compute the increment of quality by adding v_i;
    if $increment > max$ **then**
      $v = v_i$;
      Update max;
    end
  end
  $T = T \cup \{v\}$;
  Update the values $R(T, v_{i'}, a_j)$ and $P(T, l)$;
end
return T;

Algorithm 1: Approximate algorithm for S^3 model.

$R(T, v_i, a_{jl})$: representative degree of T for v_i on attribute a_j
$R(T, v, a)$ can be simply defined as if some user from T has the same value of attribute a as v, then $R(T, v, a) = 1$; otherwise $R(T, v, a) = 0$.

$$P(T, l) = \sum_{v_i \in V_l} R(T, v_i, a_{j_l})$$

$$Q(G, X, \mathbf{G}, T) = \sum_{(V_l, a_{j_l}) \in \mathbf{G}} m_l \{P(T, l)\}^{\beta}$$

Theoretical Analysis

Input: The set of users V; the set of attributes A; the attributes values X; the values $R(T, v_i, a_j)$ for every $|T| = 1$ and $v_i \in V$, $a_j \in A$; the number of users to find k.
Output: A set T of k representative users.

```
T = ∅;
while |T| < k do
    max = −1;
    foreach v_i ∉ T do
        foreach (V_l, a_{j_l}) ∈ G that v_i can contribute do
            foreach v_{i'} ∈ V_l that R({v_i}, v_{i'}, a_{j_l}) > 0 do
                Compute the increment of R(T, v_{i'}, a_{j_l});
            end
            Compute the total increment of P(T, l);
        end
        Compute the increment of quality by adding v_i;
        if increment > max then
            v = v_i;
            Update max;
        end
    end
    T = T ∪ {v};
    Update the values R(T, v_{i'}, a_j) and P(T, l);
end
return T;
```

Algorithm 1: Approximate algorithm for S^3 model.

- The greedy algorithm for the S3 model can guarantee a (1-1/e) approximation

$$Q(G, X, \mathbf{G}, T) = \sum_{(V_l, a_{j_l}) \in \mathbf{G}} m_l \{P(T, l)\}^{\beta}$$

Submodular property

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$$

Let $\Delta_k = f(T^*) - f(T^k)$

where T^* is the optimal solution;
T^k is the solution obtained in the *k-th* iteration.

Thus $\Delta_k \leq k(\Delta_j - \Delta_{j+1}), \quad 0 \leq j \leq k-1$

$$\Delta_k \leq \left(1 - \frac{1}{k}\right)^k \Delta_0 \leq \frac{1}{e} f(T^*)$$

Recall $e^x \geq 1 + x$

Finally $f(T^k) \geq \left(1 - \frac{1}{e}\right) f(T^*)$

Strategic Sampling for Diversity (SSD)

- Maximize the diversification of the selected representative users

$$Q(G, X, \mathbf{G}, T) = \min_{(V_l, a_{j_l}) \in \mathbf{G}} \{\lambda_l \cdot P(T, l)\}$$

Diversity parameter
We set it as $\lambda_l = 1/|V_l|$

- Approximation Algorithm
  - each time we choose the poorest group (with the smallest $\lambda_l \bullet P(T, l)$),
  - and then select the user who maximally increases the representative degree of this group

Theoretical Analysis

- **Theorem 2.** *Suppose the representative set generated by S3 is T, the optimal set is T*. Then the greedy algorithm for SSD can guarantee an approximation ratio of C/d•Q*, where d is the number of attributes, C is a constant, and Q* is the optimal solution.*
- Proof. The proof is based on the proof for S3.

Experiments

Datasets

- Co-author network: ArnetMiner (http://aminer.org)
  - Network: Authors and coauthor relationships from major conferences
  - Attributes: Keywords from the papers in each dataset as attributes
  - Ground truth: Take program committee (PC) members of the conferences during 2007-2009 as representative users

- Microblog network: Sina Weibo (http://weibo.com)
  - Network: users and following relationships
  - Attributes: location, gender, registration date, verified type, status, description and the number of friends or followers
  - No ground truth

Datasets: Co-author and Weibo

Coauthor

Dataset	Conf.	#nodes	#edges	#PC
Database	SIGMOD	3,447	9,507	256
	VLDB	3,606	8,943	251
	ICE	4,150	9,120	244
SUM	ALL	8,027	23,770	291
Data Mining	SIGKDD	2,494	4,898	243
	ICDM	2,121	3,452	211
	CIKM	2,942	4,921	205
SUM	ALL	6,394	12,454	373

Weibo

Dataset	#nodes	#edges
Program	19,152	19,225
Food	189,176	204,863
Student	79,052	76,559
Public welfare	324,594	383,702

Accuracy Performance

Table 2: Performance of different methods on the Database dataset (%). (In Model 1, we set $\beta = 0.6$.)

Methods	P@10	P@50	P@100	R@50	R@100	F1
InDegree	50.0	36.0	40.0	6.2	13.7	33.0
HITS_h	40.0	30.0	27.0	5.2	9.3	20.3
HITS_a	30.0	30.0	30.0	5.2	10.3	24.1
PageRank	40.0	26.0	27.0	4.5	9.3	24.7
S^3	70.0	46.0	46.0	7.9	15.8	41.9
SSD	10.0	36.0	39.0	6.2	13.4	32.6

Results: S3 outperforms all the other methods In terms of P@10, P@50 and achieves the best F1 score

Comparison methods:

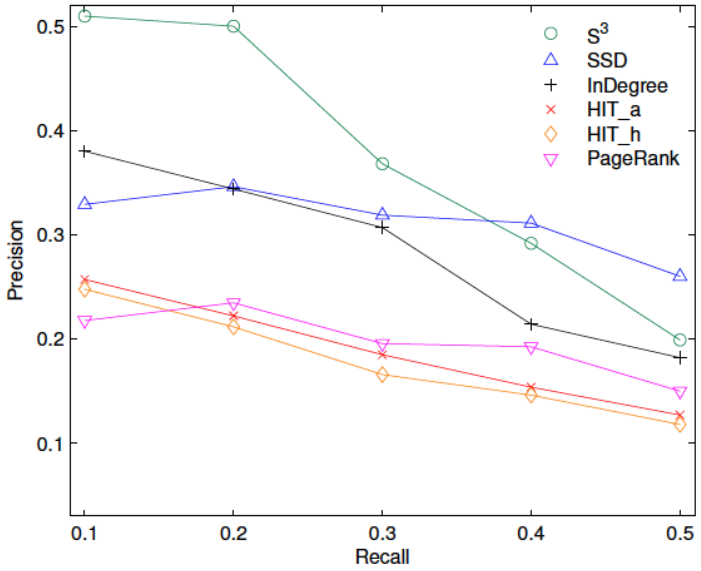
- **InDegree:** select representative users by the number of indegree
- **HITS_h and HITS_a:** first apply HITS algorithm to obtain authority and hub scores of each node. Then the two methods respectively select representative users according to the two scores
- **PageRank:** select representative users according to the pagerank score

Table 3: Performance of different methods on the Data Mining dataset (%). (In Model 1, we set $\beta = 0.6$.)

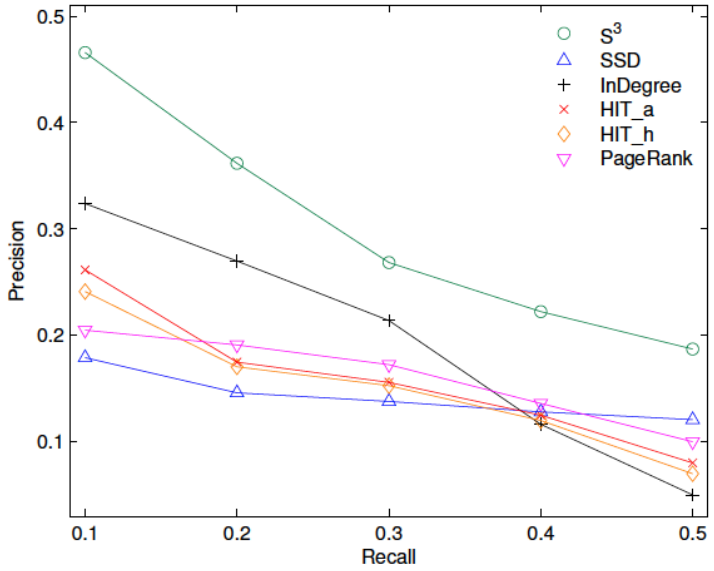
Methods	P@10	P@50	P@100	R@50	R@100	F1
InDegree	80.0	62.0	45.0	8.3	12.1	28.4
HITS_h	70.0	42.0	30.0	5.6	8.0	21.4
HITS_a	70.0	48.0	30.0	6.4	8.0	18.8
PageRank	50.0	38.0	30.0	5.1	8.0	24.9
S^3	80.0	64.0	53.0	8.6	14.2	36.7
SSD	20.0	32.0	22.0	4.3	5.9	22.5

Accuracy Performance

- The precision-recall curve of the different methods



(a) Database



(b) Data Mining

Efficiency Performance

Table 4: Efficiency performance of different methods on the three datasets. The CPU time does not include the time of loading the network data. (Second)

Methods	Database (291)	Data Mining (373)	Weibo (200)	
			Program	Food
InDegree	0.10	0.16	0.01	0.02
HITS_h	2.44	1.58	0.99	45.66
HITS_a	2.44	1.58	0.95	45.60
PageRank	6.89	2.95	0.51	18.33
S^3	1.68	3.62	4.45	50.71
SSD	0.54	0.74	0.67	10.81

S3 and SSD respectively only need 50 and 10 seconds to perform the sampling over a network of ~200,000 nodes

Case Study

Database		Data Mining	
S3	PageRank	S3	PageRank
Jiawei Han	Serge Abiteboul	Philip S. Yu	Philip S. Yu
Jeffrey F. Naughton	Rakesh Agrawal	Jiawei Han	Jiawei Han
Beng Chin Ooi	Michael J. Carey	Christos Faloutsos	Jian Pei
Samuel Madden	Jiawei Han	ChengXiang Zhai	Christos Faloutsos
Johannes Gehrke	Michael Stonebraker	Bing Liu	Ke Wang
Kian-Lee Tan	Manish Bhide	Vipin Kumar	Wei-Ying Ma
Surajit Chaudhuri	Ajay Gupta	Jieping Ye	Jianyong Wang
Elke A. Rundensteiner	H. V. Jagadish	Ming-Syan Chen	Jeffrey Xu Yu
Divyakant Agrawal	Surajit Chaudhuri	Padhraic Smyth	Haixun Wang
Wei Wang	Warren Shen	C. Lee Giles	Hongjun Lu

Advantages of S3:
(1) S3 tends to select users with more diverse attributes;
(2) S3 can tell on which attribute the selected users can represent.

Conclusions

- Formulate a novel problem of *Q-evaluated Sampling Representative Users* (Q-SRU)
- Theoretically prove the NP-Hardness of the Q-SRU problem
- Propose two instantiation sampling models
- Present efficient algorithms with provable approximation guarantees

Sampling Representative Users from Large Social Networks

Jie Tang, Chenhui Zhang

Tsinghua University

Keke Cai, Li Zhang, Zhong Su

IBM, China Research Lab

Download Code&Data here: http://aminer.org/repuser

Theoretical Analysis

$T\downarrow l$: the people in T that contributes to attribute group $(G\downarrow l, a\downarrow j\downarrow l)$
$T\downarrow l\uparrow *$: the people in $T\uparrow *$ that contributes to attribute group $(G\downarrow l, a\downarrow j\downarrow l)$

For a fixed attribute group $(G\downarrow l, a\downarrow j\downarrow l) \in \boldsymbol{G}$, the function $P(T,l)$ is submodular

Let be maximal number such that the first elements selected by greedy are all contained in

$$\sum_{(G_l, a_{j_l}) \in \mathbf{G}} w_l \geq k \quad (1)$$

If $w_l \geq |T_l^*|$ there must be $P(T,l) \geq (1-1/e)P(T_l^*, l) = (1-1/e)P(T^*, l)$

If $w_l < |T_l^*|$

we can continue selecting people for $(G\downarrow l, a\downarrow j\downarrow l)$ until $|T\downarrow l\uparrow *|$ people are selected. And the set becomes T'

$$P(T, G_l) \geq \frac{w_l}{|T_l^*|} \cdot P(T', G_l) \geq \frac{w_l}{|T_l^*|} \cdot (1-1/e)P(T^*, G_l)$$

Theoretical Analysis

$$P(T,l) \geq C/d \cdot P(T^*,l) \quad (2)$$

Therefore by inequality (2)

$$w_{l_0} < \lfloor |T^*_{l_0}|/d \rfloor \quad (3)$$

We argue that for any $l \neq l \downarrow 0$, there must be

$$w_l \leq \lfloor |T^*_l|/d \rfloor \quad (4)$$

By inequalities (3) and (4), the summation

$$\sum_{(G_l,a_{j_l})\in\mathbf{G}} w_l < \sum_{(G_l,a_{j_l})\in\mathbf{G}} \lfloor |T^*_l|/d \rfloor \leq \sum_{(G_l,a_{j_l})\in\mathbf{G}} |T^*_l|/d \leq k$$

As we get a contradiction with inequality (1), this theorem is thus proved.