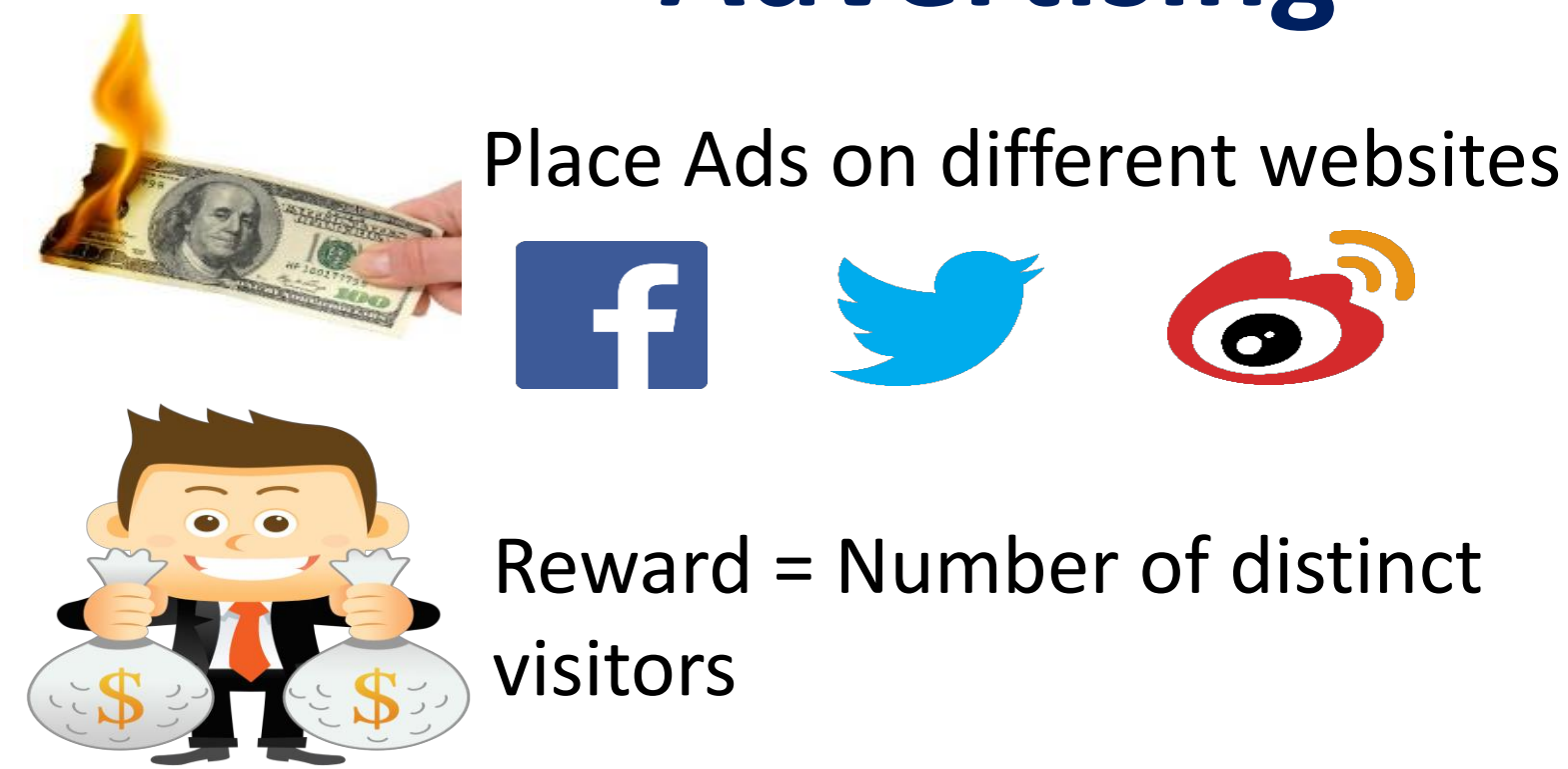




Motivating Example: Online Advertising



Goal: Maximize the reward (# distinct visitors)

Challenges

- **Limited Budget.** How to allocate the budget on each website?
- **Duplicate Visitors.** The same visitor may visit the website several times.
- **Unknown Possible Visitor Numbers.** How to learn the parameters?

Model Descriptions

- ❖ **Communities:** **disjoint** sets C_1, \dots, C_m
 - Community sizes $d = \{d_1, \dots, d_m\}$
 - Example: Visitors of different websites
- ❖ **Explore Community C_i**
 - Explore once, **meet** a member in C_i *uniformly at random*
- ❖ **Reward:** # of *distinct* members $\in C_1 \cup \dots \cup C_m$
- ❖ **Budget:** explore communities at most K times



Our Results

Offline Problems (Non-adaptive, and adaptive exploration)

- ⚙ **Setting:** Community sizes are **known**
- 🔑 **Solution:** Greedy method/policy
- 📦 **Conclusion:** Greedy method/policy is optimal

Online Learning (Non-adaptive, and adaptive exploration)

- ⚙ **Setting:** Community sizes are **unknown**
- 🔑 **Solution:** Combinatorial Lower Confidence Bound (CLCB) algorithm
- 📦 **Conclusion:** Logarithmic/constant regret bound

Offline Optimization

Key Assumption: The sizes of communities $d = \{d_1, \dots, d_m\}$ are **known**



Problem 1. Offline Non-adaptive

Non-adaptive Exploration

- Determine the budget allocation $k = \{k_1, \dots, k_m\}$ before the exploration
- Explore C_i for k_i times
- $\sum_{i \in [m]} k_i = K$

Goal. Find an **optimal budget allocation** $k^* = \{k_1^*, \dots, k_m^*\}$ to **maximize** the expected reward

$$\max_k \sum_{i=1}^m d_i \left(1 - \left(1 - \frac{1}{d_i}\right)^{k_i}\right), \quad s.t. \sum_{i \in [m]} k_i = K$$

 **Expected Reward**  **Constraint**

Greedy Method (time complexity $O(m \log m)$)

Start from $k_i = \left\lfloor \frac{(K-m) \ln(1-1/d_i)}{\sum_{j=1}^m \ln(1-1/d_j)} \right\rfloor$ (good initial point)

At each step (**stop** when $\sum_{i \in [m]} k_i = K$), we choose community C_{i^*} such that

$$i^* \in \arg \max_{i \in [m]} \left(1 - \frac{1}{d_i}\right)^{k_i}, \quad k_{i^*} \leftarrow k_{i^*} + 1$$

Theorem: Greedy method obtains **optimal** budget allocation!

Problem 2. Offline Adaptive

Adaptive Exploration

- Step by step exploration
- Choose community to explore based on previous results

Goal. Find an **optimal policy** π^* to **maximize** the expected reward

- Mapping function $\pi(\text{previous results}) = \text{next community to explore}$

Greedy Policy π^g

- At each step, choose the community which has the largest percentage of unvisited members

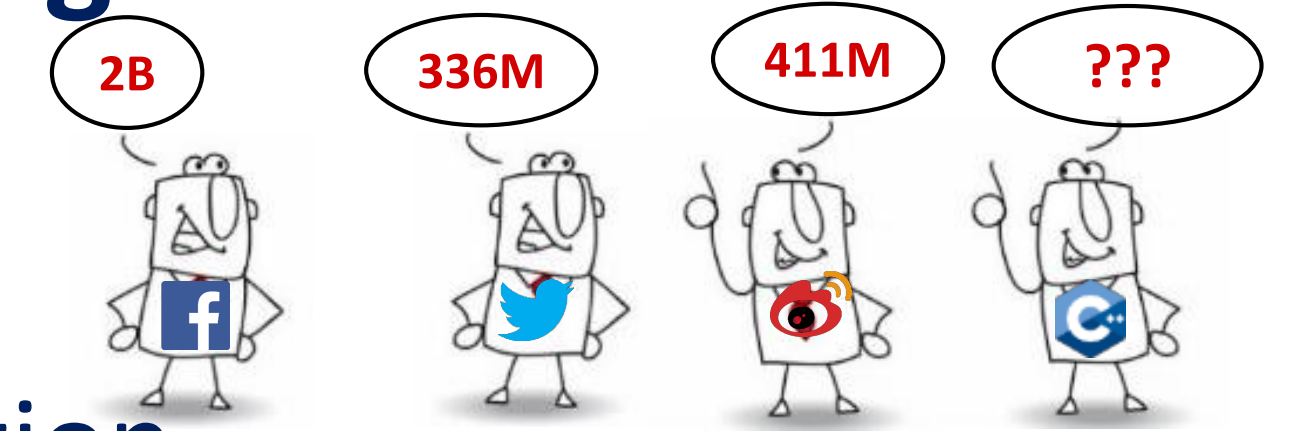
Adaptive submodular $\longrightarrow 1 - 1/e?$ ❌

Theorem: Greedy policy is the **optimal** among all the policies!

- Proved by inductive reasoning
- Applied to the general reward function $f(\# \text{distinct members})$

Online Learning

- The sizes of communities are **unknown**
- Online advertising in **multiple rounds**
 - In each round, we try to maximize the reward



Problem Definition

Problem 3. Online Non-adaptive

For each round $t = 1, \dots, T$

- Choose "action" $k_t = (k_{1,t}, \dots, k_{m,t})$ based on previous exploration results
- Explore community C_i for $k_{i,t}$ times (*non-adaptive exploration*)

Problem 4. Online Adaptive

For each round $t = 1, \dots, T$

- Choose an "action" π_t based on previous exploration results
- Explore community with policy π_t (*adaptive exploration*)

Goal. **Maximize** the cumulative rewards in T rounds

Bandit Algorithm

At round t ($\hat{\mu}_{i,t}$: unbiased estimator of $1/d_i$)

Compute lower confidence bound

- Radius $\rho_{i,t} \leftarrow \sqrt{\frac{3 \ln t}{2T_{i,t-1}}}$
- $\underline{\mu}_{i,t} \leftarrow \max\{0, \hat{\mu}_{i,t} - \rho_{i,t}\}$

Combinatorial Lower Confidence Bound (CLCB)

Combinatorial Multi-Armed Bandit: General Framework, Results and Applications

Play Oracle $\left(\left(\frac{1}{\underline{\mu}_{1,t}}, \dots, \frac{1}{\underline{\mu}_{m,t}}\right)\right)$

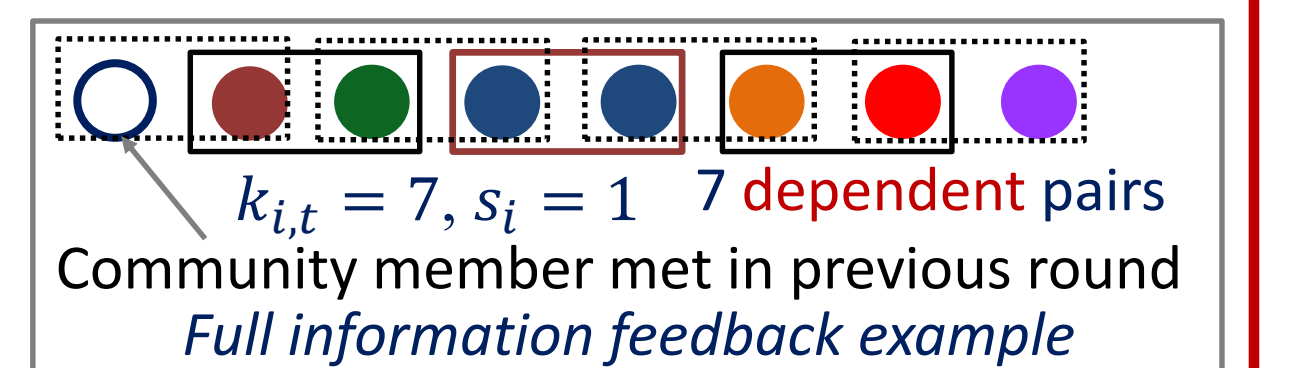
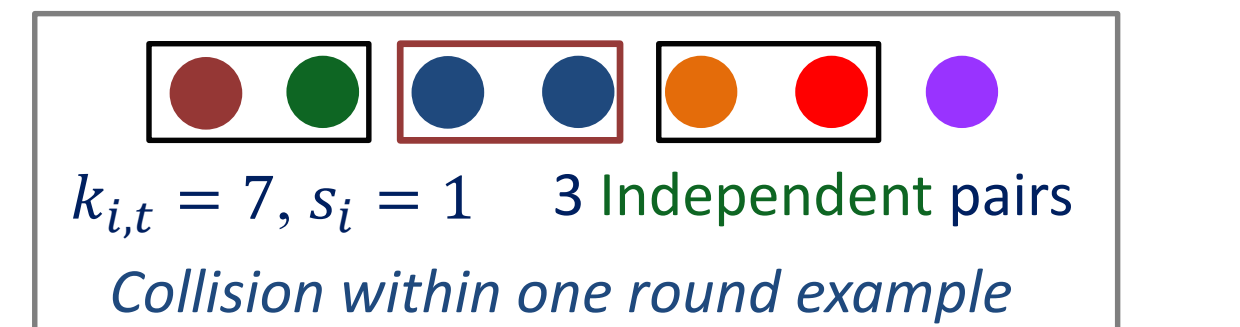
- $S_{i,t}$: set of met members in community C_i in round t

Online, Non-adaptive

- Start from $k_i = 0$
- At each step, choose $i^* \in \arg \max_i (1 - \underline{\mu}_{i,t})^{k_i}$
- $k_{i^*} \leftarrow k_{i^*} + 1$

Online, Adaptive

- At each step, choose $i^* \in \arg \max_i 1 - \frac{\mu_{i,t}}{c_i}$
- c_i : # members that is already met in C_i



Update estimates

- $k_{i,t}: |S_{i,t}|$
- Collision within one round
 - $s_i = \sum_{x=1}^{\lfloor k_{i,t}/2 \rfloor} \mathbb{I}\{S_i[2x-1] = S_i[2x]\}, T_{i,t} \leftarrow T_{i,t-1} + \lfloor k_{i,t}/2 \rfloor$
- Full information feedback
 - $s_i = \sum_{x=1}^{k_{i,t}} \mathbb{I}\{S_i[x-1] = S_i[x]\}, T_{i,t} \leftarrow T_{i,t-1} + k_{i,t}$
- $S_{i,t} \leftarrow S_{i,t-1} \cup s_i, \hat{\mu}_{i,t} = S_{i,t}/T_{i,t}$

Regret Bound

Regret Bound (Problem 3&4)

- Collisions within one round: **Reg(T) ~ O(log T)** E.g., non-adaptive regret $O\left(\frac{(K-m+1)^3 \log T}{\Delta_{\min}^i}\right)$
 - Tighter bound, leverage existing analysis framework
- Full information feedback: **problem dependent const. O(1)**