INFORMATION CASCADE AT GROUP SCALE

Milad Eftekhar, Yashar Ganjali, Nick Koudas

University of Toronto

Introduction

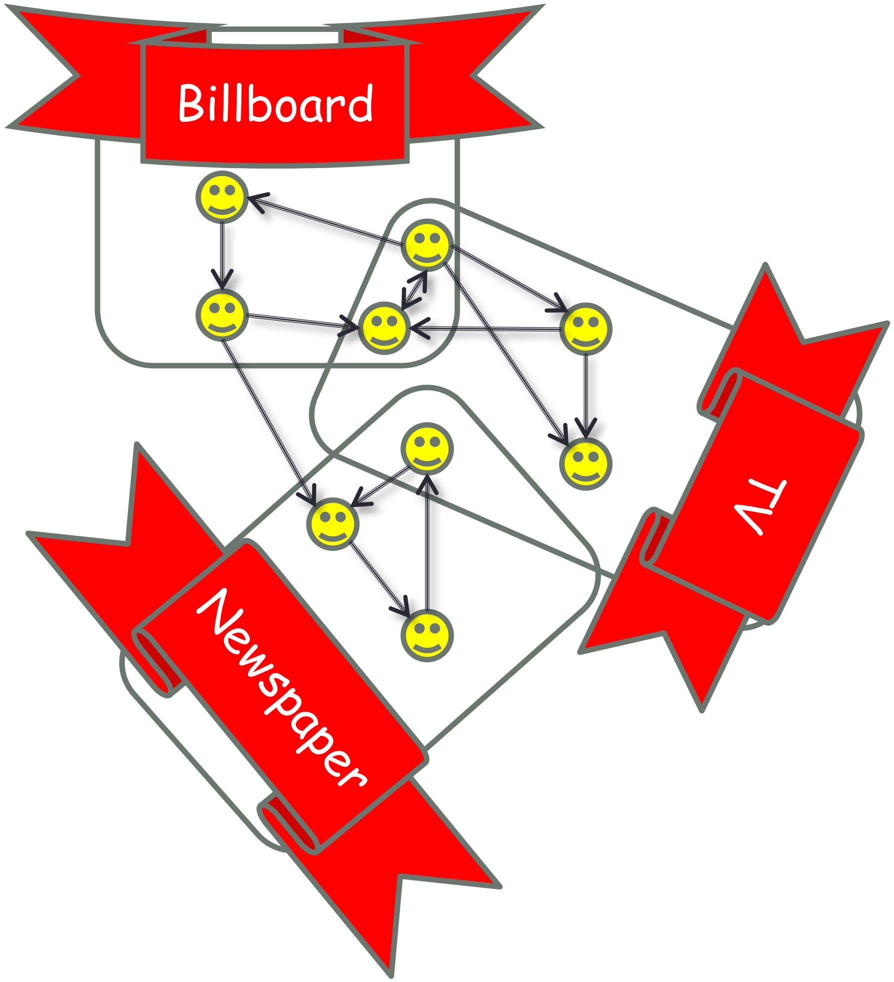
- Identifying the k most influential **individuals** is a well-studied problem.

- We **generalize** this problem to identify the l most influential **groups**.

- Application:
  - Companies often target groups of people
    - E.g. by billboards, TV commercials, newspaper ads, etc.

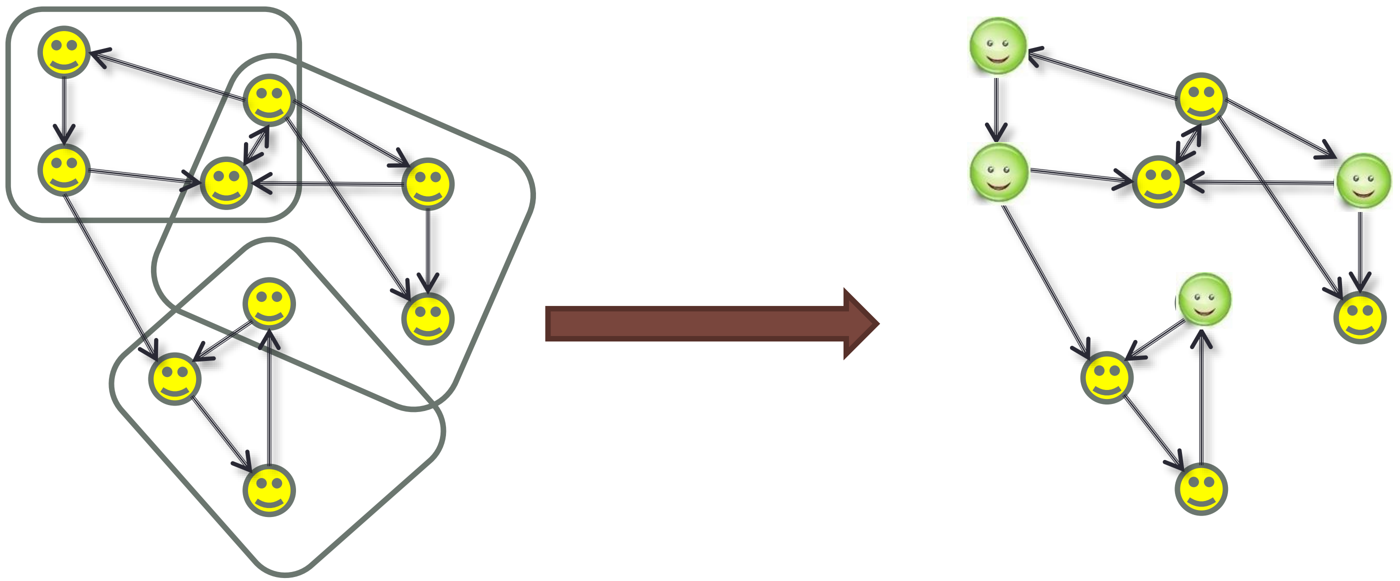
Group targeting

- Groups
- Advantages
  - Improved performance
  - Natural targets for advertising
  - An economical choice



Fine-Grained Diffusion (FGD)

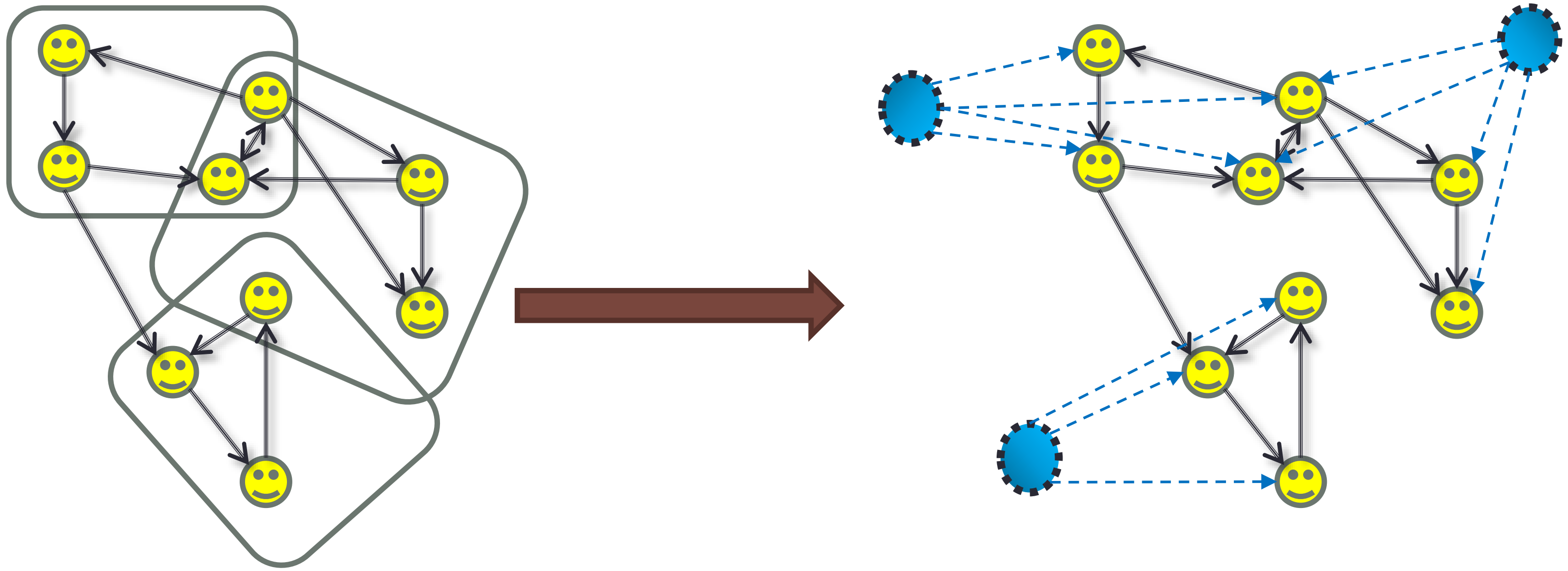
- Determine how advertising to a group translates into individual adopters.



- Run individual diffusion process on these adopters.

FGD Modeling

- Graph G': add a node for each group, add edges between a node corresponding to a group g_i and its members with weight w_i that depends on
  - Advertising budget, size of group, the escalation factor, and the budget needed to convince an individual



FGD Modeling (Cont'd)

- Escalation Factor β: how many more initial adaptors we can get by group targeting rather than individual targeting.

Individual Targeting	Group Targeting
Advertising budget $1000	Advertising budget $1000
Cost of convincing an individual $100	Billboard Cost = $1000 Audience = 10000 individuals
10 initial adopters	20 initial adopters

$$\beta = \frac{20}{10} = 2$$

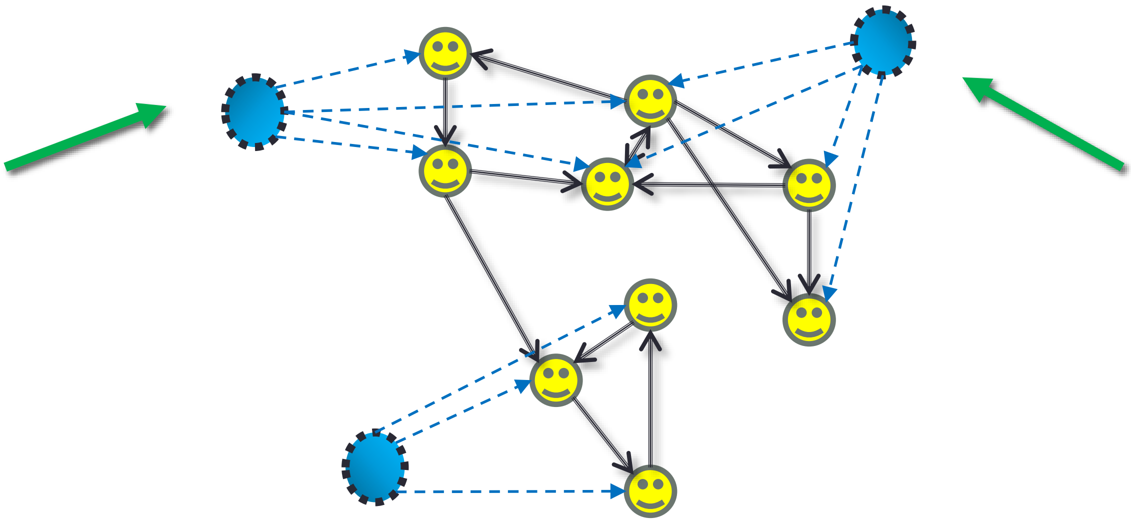
FGD Modeling (Cont'd)

- Escalation Factor β
  - Based on the problem structure, the size and shape of the network, the initial advertising method, etc.
    - Individual advertising: $\beta = 1$
    - Billboard advertising: $\beta = 200$
    - Online advertising: $\beta = 400$

Problem statement

- **Goal**: Find the l most influential groups (blue group-nodes)

Top-2 influential groups



Top-2 influential groups

- NP-hard under FGD model

topfgd algorithm

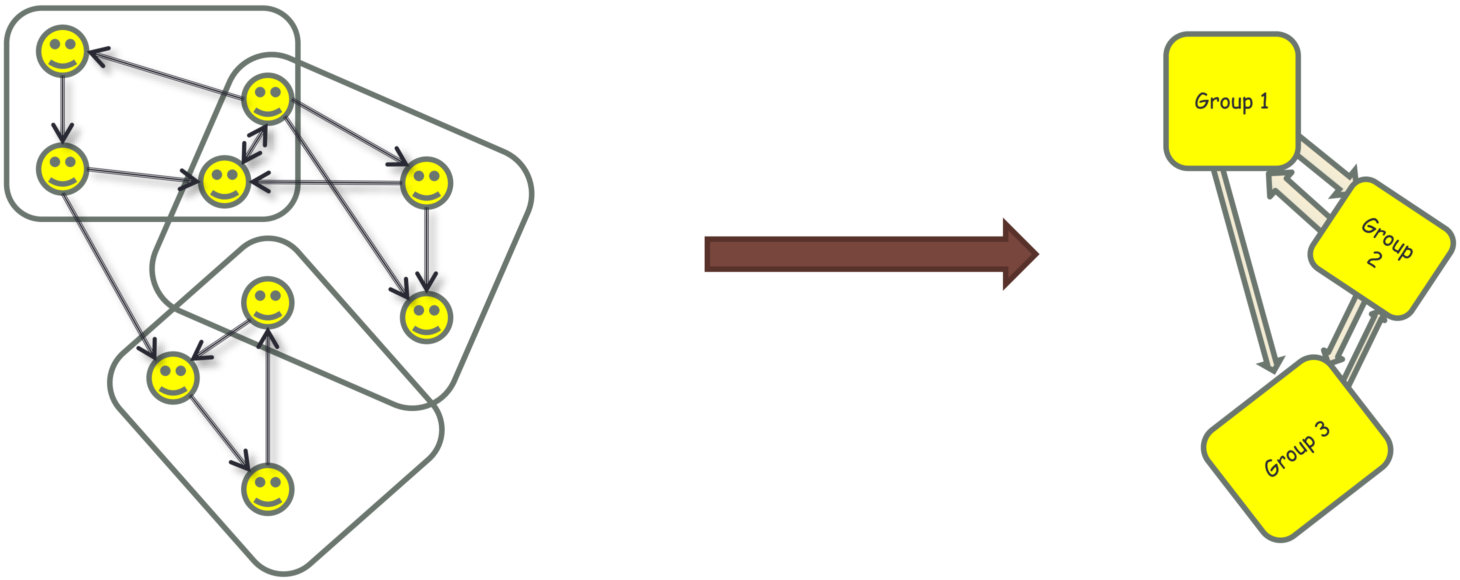
- Diffusion in FGD is monotone and submodular

- topfgd: a greedy algorithm provides a (1-1/e) approximation factor.
  - In each iteration, add the group resulting to the maximum marginal increase in the final influence.

- Time: $O(l \times m \times |E_{ind}| \times R)$

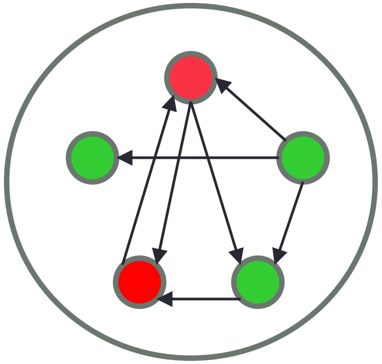
Coarse-Graind Diffusion (CGD)

- FGD is not practical for large social networks
- Idea: incorporate information about individuals without running explicitly on the level of individuals
  - A graph to model inter-group influences



CGD Modeling

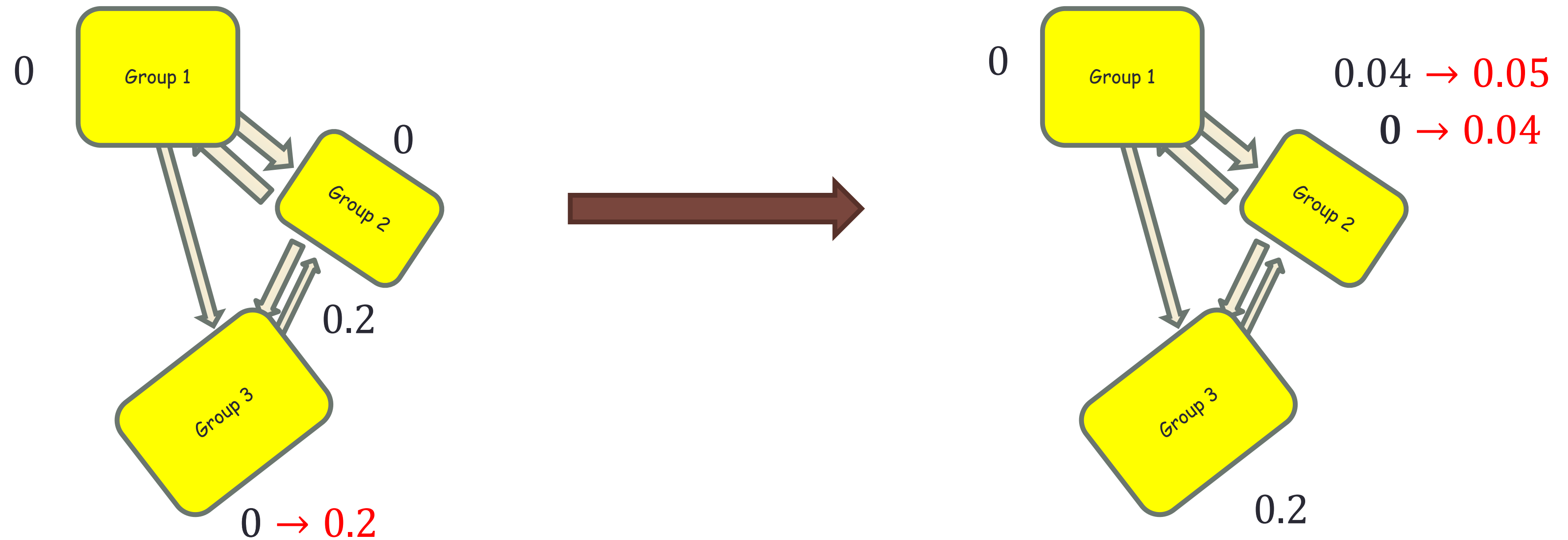
- Differences with “Individual Diffusion” models
  - No binary decisions
    - **Progress fraction** for each group
  - Two types of diffusion
    - Inter-group diffusion
    - Intra-group diffusion
  - Submodularity?



Progress fraction = 0.6

CGD Diffusion Model

- Each newly activated fraction of a group can activate its neighboring groups
  - As a result of an activation attempt from A to B, some activation attempts also occur between members of B
- Continue for several iterations to converge



topcgd algorithm

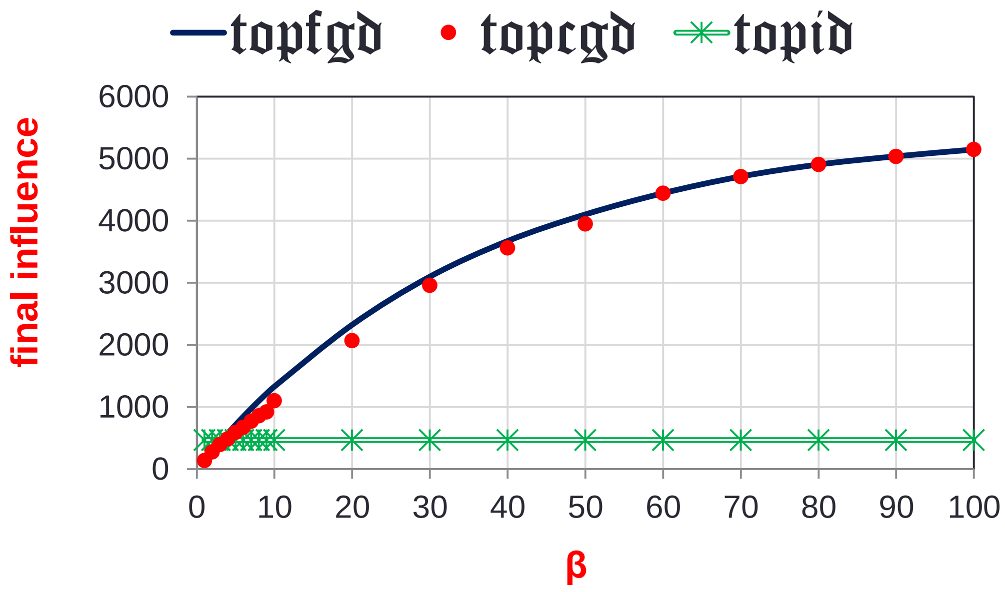
- **Goal**: Find the l most influential groups
  - NP-hard under CGD model
- Diffusion in CGD is monotone and submodular
- **topcgd**: a greedy algorithm provides a (1-1/e) approximation factor.
- Time: $O(|E_{ind}| + ml(mt + n))$
  - t is the number of iterations to converge (~10)

Experimental setup

- Datasets:
  - DBLP: 800K nodes, 6.3M edges, 3200 groups
- Comparison
  - Spend same advertising budget on all algorithms
  - Measure the final influence (the number of convinced individuals)
    - Run Individual Diffusion process on the initial convinced individuals

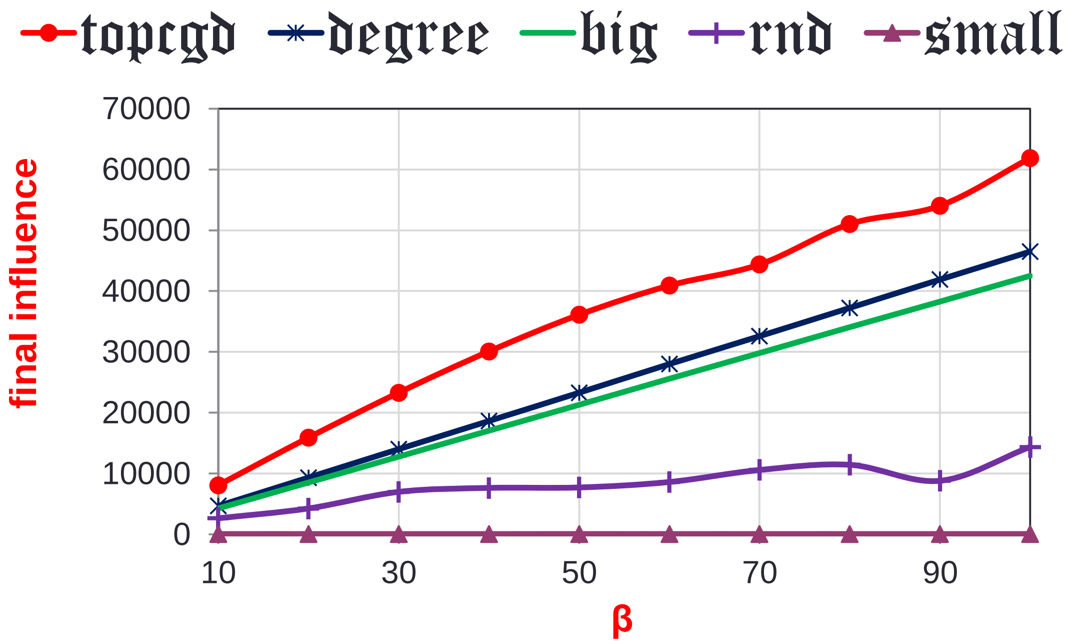
Results

- DBLP-1980: 8000 nodes, 69 groups
- Compare $\mathfrak{topid}$ vs. $\mathfrak{topfgd}$ vs. $\mathfrak{topcgd}$
- Final influence: $\mathfrak{topfgd}$ and $\mathfrak{topcgd}$ outperform $\mathfrak{topid}$ for $\beta > 3$
- Time: $\mathfrak{topid}$ (30 days), $\mathfrak{topfgd}$ (an hour), $\mathfrak{topcgd}$ (0.2 sec)



Results (Cont'd)

- DBLP: topcgd vs. Baselines
  - rnd, small, big, degree
- Time of topcgd: 100 minutes
- topfgd and topid not practical



Conclusion and Future Works

- Focus on groups rather than individuals
  - Wider diffusion
  - Improved performance
  - More less influential individuals vs. less more influential individuals
- Although CGD aggregates the information about individuals (hence improved performance), it results to final influence comparable to FGD.

- We are interested in a generalized model where
  - Groups are allowed to receive different budgets
  - The cost of advertising to each group is predetermined

Thanks!

(Questions?)