香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Efficient Influence Maximization in Social Networks

Presented by WAN, Pengfei
Dept. ECE, HKUST

Wei Chen, et al, “Efficient Influence Maximization in Social Networks”, KDD09’

OUTLINE

- Problem
- Previous Work
- Degree Discount Heuristics
- Summary
- References

Problem Statement

- Find a small subset of nodes in a social network that could maximize the spread of influences.
- Known as *Influence Maximization*
- A.k.a *Viral Marketing* which makes use of "word-of-mouth marketing" properties of social network

Problem Statement

- Optimization problem first introduced by Domingos and Rechardson, KDD01'/02', *NP-hard to solve*

- Elegant graph formulation introduced by Kempe, et al, KDD03'

Given:

- ✓ A graph G(V, E):
  - --Vertices: individuals in social network
  - --Edges: connection or relationship
- ✓ k, size of output seeds
- ✓ A cascade model: LTM, ICM

Output:

S, a set of seeds (nodes) that maximize the expected number of nodes active in the end

Problem Statement: Cascade Model

- Models how influences propagate
- Linear Threshold Model *(LTM)*
- Independent Cascade Model *(ICM)*
- … …
- Analogous to Epidemic Models like SIS, SIR

Linear Threshold Model

- A node *u* has random threshold $\theta_u \sim U[0,1]$

- A node *u* is influenced by each neighbor *v* according to a *weight* b_{uv} witch satisfies:

$$\sum_{\text{v neighbor of u}} b_{u,v} = 1$$

- A node *u* becomes active when at least θ_u fraction of its neighbors are active

$$\sum_{\text{v active neighbor of u}} b_{u,v} \geq \theta_u$$

Independent Cascade Model

- When node *u* becomes active, it has a *single* chance of activating each currently inactive neighbor *v.*

- The activation attempt succeeds with probability p_{uv} .

- In both LTM and ICM, active nodes never deactivate.

OUTLINE

- Problem
- Previous Work
- Degree Discount Heuristics
- Summary
- References

Previous Work: “Maximizing the Spread of Influence Through a Social Network”,KDD03’

- Proposed by D.Kempe, J.Kleinberg and E.Tardos

- Greedy hill-climbing algorithm:

  In each round add a vertex v^* into S such that v^* and S maximize the influence spread f:

$$v^* = arg \max_{v} f(S+v) - f(S)$$

- Monte Carlo:

  Influence spread is estimated with R repeated simulations

- Effectiveness:

  Can guarantees a solution with (1 – 1/e) of the optimal

- Drawback:

  poor efficiency, 15,000 nodes takes a few days to compete

Previous Work:
"Cost-effective Outbreak Detection in Networks", KDD07'

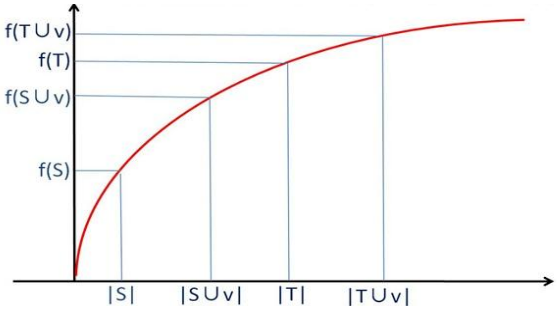
- Proposed by J. Leskovec, A. Krause, et al

- Cost-effective Lazy Forward algorithm:

The CELF optimization utilizes submodularity of influence spread function to greatly reduce the number of evaluations of vertices, and get the same performance as the original greedy algorithm.

- Submodularity:

$$\forall S \subset T \subset N, \forall v \in N \setminus T,$$
$$f(S+v) - f(S) \geq f(T+v) - f(T)$$



- Efficiency:

approximately 700 times fast than original greedy algorithm, but still hours to finish.

OUTLINE

- Problem
- Previous Work
- Degree Discount Heuristics
- Summary
- References

Degree Discount Heuristics

- Proposed by W.Chen, Y.Wang , S.Yang from MSRA and Tsinghua

- High Efficiency:

  Amazingly reduces the running time by over *six orders* of magnitude with *less than 3.5%* degradation in performance.

- Motivation:

  Conventional degree/centrality based heuristics perform poorly in practical scenarios because they *ignore the network effect*.

  Important Fact: Since many of the most central nodes may be clustered, targeting all of them is not at all necessary.

Degree Discount Heuristics

- **Basic Idea**

Consider edge $\overline{uv}$, with *u* in the seed set S and *v* being considered. Since *u* is in the seed set, by taking network effect into consideration, we should not count edge $\overline{uv}$ towards v's degree. i.e. Degree Discount

- **Assumption**

In ICM, when propagation probability *p* is small, we may ignore indirect influence of *v* to multi-hop neighbors and focus on the *direct influence* of *v* to its immediate neighbors.

Remarks : Is this assumption still reasonable when k is small ? Or when neighbor overlapping is prominent?

Degree Discount Heuristics

- ## Degree Discount Model:

  t_v -- number of *v*'s neighbors that in seed set *S*

  d_v-- degree of node *v*

✓ Probability that v is influenced by its immediate neighbors: $1 - (1 - p)^{t_v}$
in such case, selecting v dose not contribute additional influence.

✓ Probability that v is not influenced by its immediate neighbors: $(1 - p)^{t_v}$
in such case, selecting v will in expectation influence $1 + (d_v - t_v) * p$ vertices.

So that the *expected number of additional vertices influenced by selecting v as seed* is:

$$[1 - (1 - p)^{t_v}] * 0 + [(1 - p)^{t_v}] * [1 + (d_v - t_v) * p]$$
$$= \left(1 - t_v * p + o(p)\right) * (1 + (d_v - t_v) * p)$$
$$\cong 1 + (d_v - 2t_v - (d_v - t_v) * t_v * p) * p \triangleq A$$

If no neighbor of *v* is selected as seed, the answer above is $1 + d_v * p \triangleq B$

Let γ be the degree discount caused by each neighbor in seed set, then

$$\gamma * t_v * \mathrm{p} = \mathrm{B} - \mathrm{A}$$
$$\gamma = 2 + (d_v - t_v) * p$$

Degree Discount Heuristics

- Algorithm:

Algorithm 4 DegreeDiscountIC(G, k)

initialize $S = \emptyset$
for each vertex v **do**
  compute its degree d_v
  $dd_v = d_v$
  initialize t_v to 0
end for
for $i = 1$ to k **do**
  select $u = \arg\max_v \{dd_v \mid v \in V \setminus S\}$
  $S = S \cup \{u\}$
  for each neighbor v of u and $v \in V \setminus S$ **do**
    $t_v = t_v + 1$
    $dd_v = d_v - 2t_v - (d_v - t_v)t_v p$
  end for
end for
output S

Degree Discount Heuristics

- Evaluations on NetHEPT:

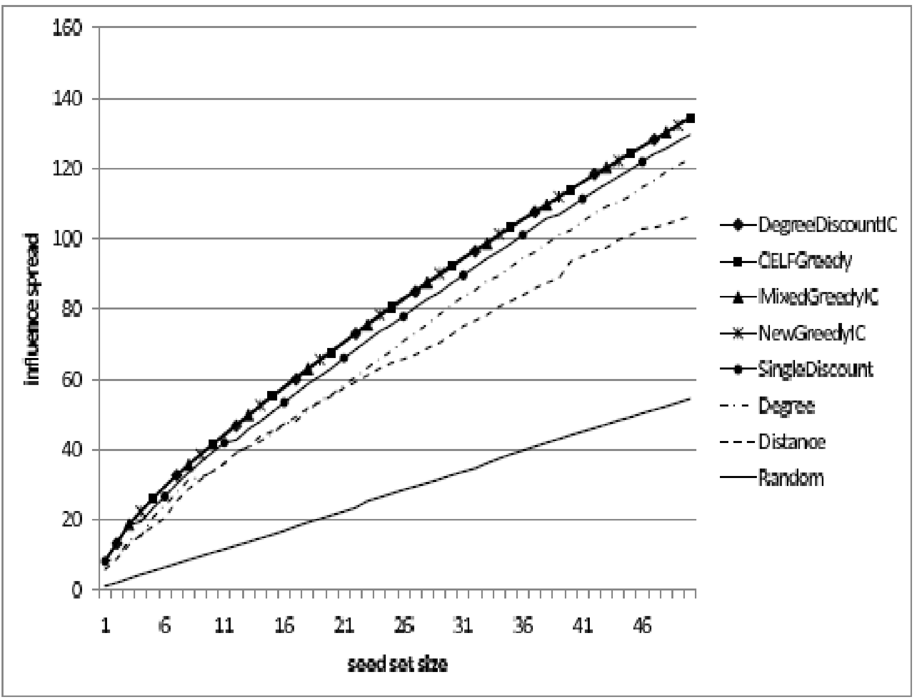


Figure 1: Influence spreads of different algorithms on the collaboration graph NetHEPT under the independent cascade model ($n = 15,233$, $m = 58,891$, and $p = 0.01$).

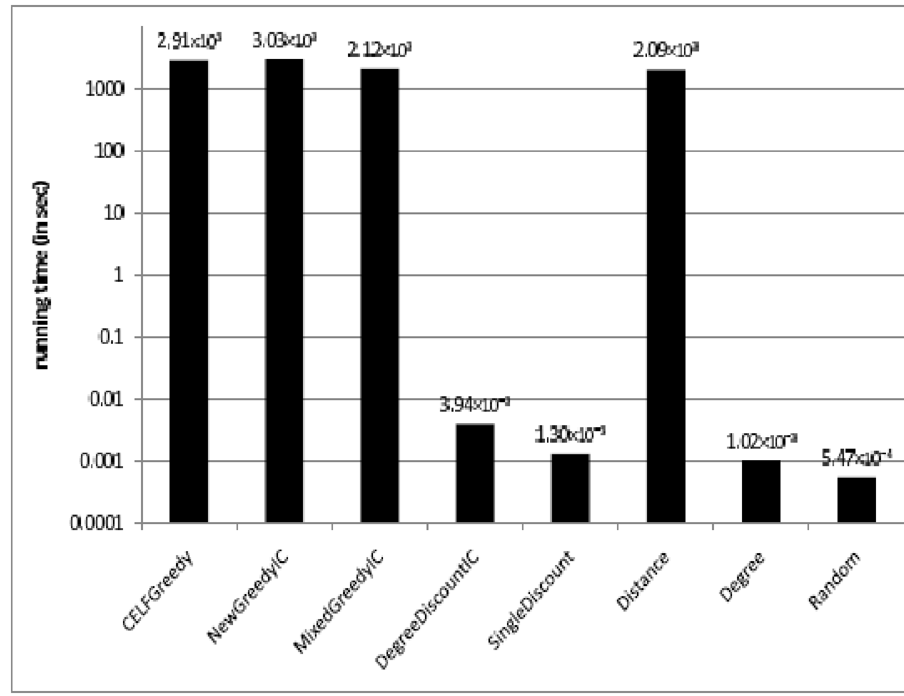


Figure 3: Running times of different algorithms on the collaboration graph NetHEPT under the independent cascade model ($n = 15,233$, $m = 58,891$, $p = 0.01$, and $k = 50$).

Degree Discount Heuristics

- Evaluations on NetPHY:

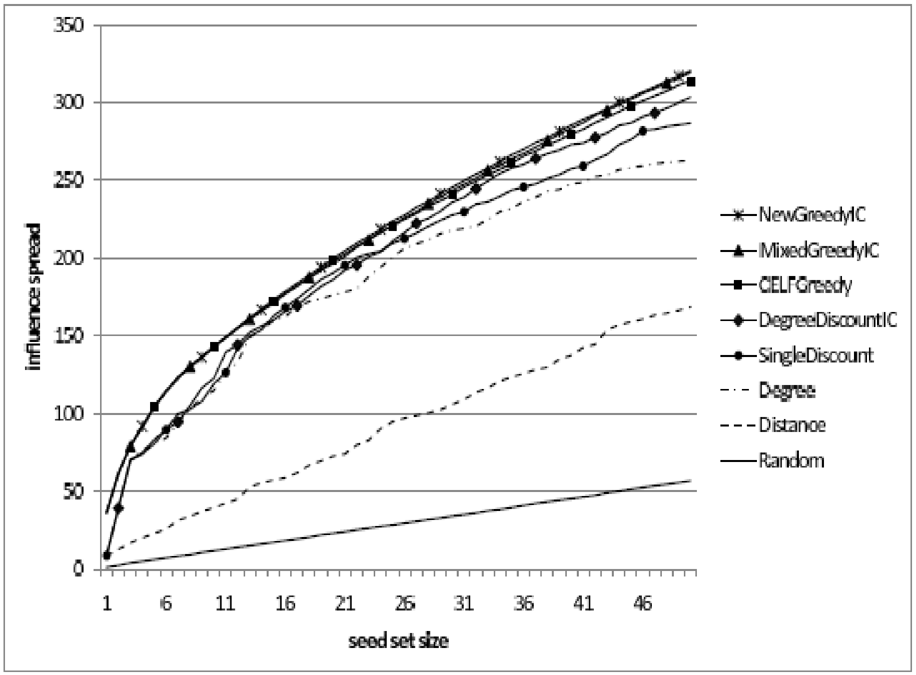


Figure 2: Influence spreads of different algorithms on the collaboration graph NetPHY under the independent cascade model ($n = 37,154$, $m = 231,584$, and $p = 0.01$).

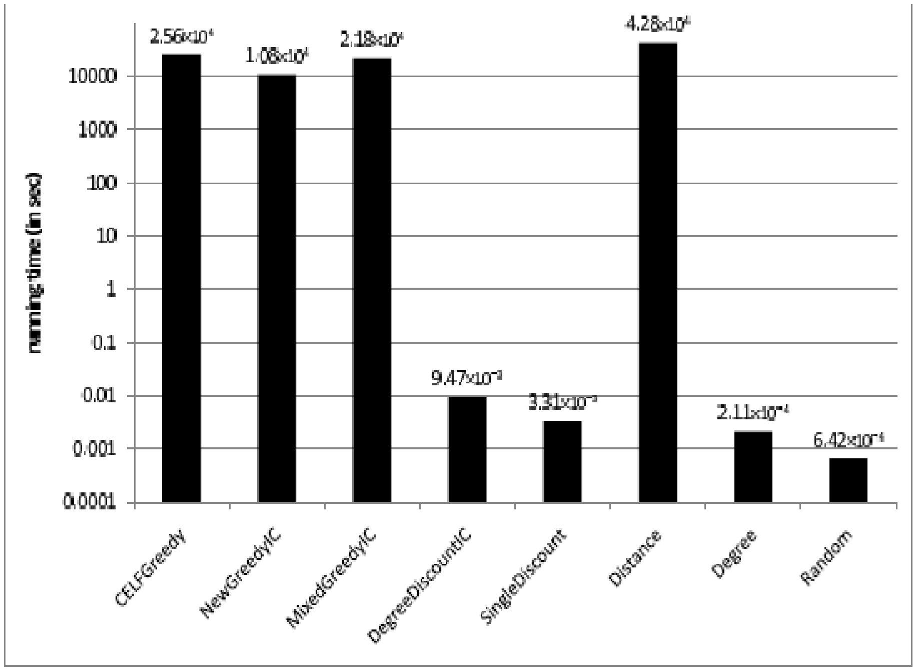


Figure 4: Running times of different algorithms on the collaboration graph NetPHY under the independent cascade model ($n = 37,154$, $m = 231,584$, $p = 0.01$, and $k = 50$).

OUTLINE

- Problem
- Previous Work
- Degree Discount Heuristics
- Summary
- References

Summary

- The current influence maximization problem is simplified, without considering other features in social networks, such as community structures and small-world phenomenon.

- The author suggests that we should focus our research efforts on searching for more effective heuristics for different influence cascade model in real life influence maximization anpplications

- More sophisticated heuristics are promising, such as taking into consideration multiple links between nodes, higher-order influences, cross-neighborhood structure…

OUTLINE

- Problem
- Previous Work
- Degree Discount Heuristics
- Summary
- References

References

- *W. Chen, Y. Wang and S. Yang ,“Efficient Influence Maximization in Social Networks”, KDD 2009*
- *D. Kempe, J. Kleinberg and E. Tardos, “Maximizing the Spread of Influence through a Social Network”, KDD 2003*

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Thank you !