Influence Maximization on Social Graphs: A Survey

Yuchen Li, Ju Fan,Member, IEEE, Yanhao Wang, and Kian-Lee Tan,Member, IEEE

Krithika Srinivasan

ks372@buffalo.edu

Introduction

What is Information Diffusion?

- A piece of information becoming pervasive through word of mouth in a social network
- Becoming increasingly relevant in the age of online social networks
- Has attracted research efforts in multiple fields

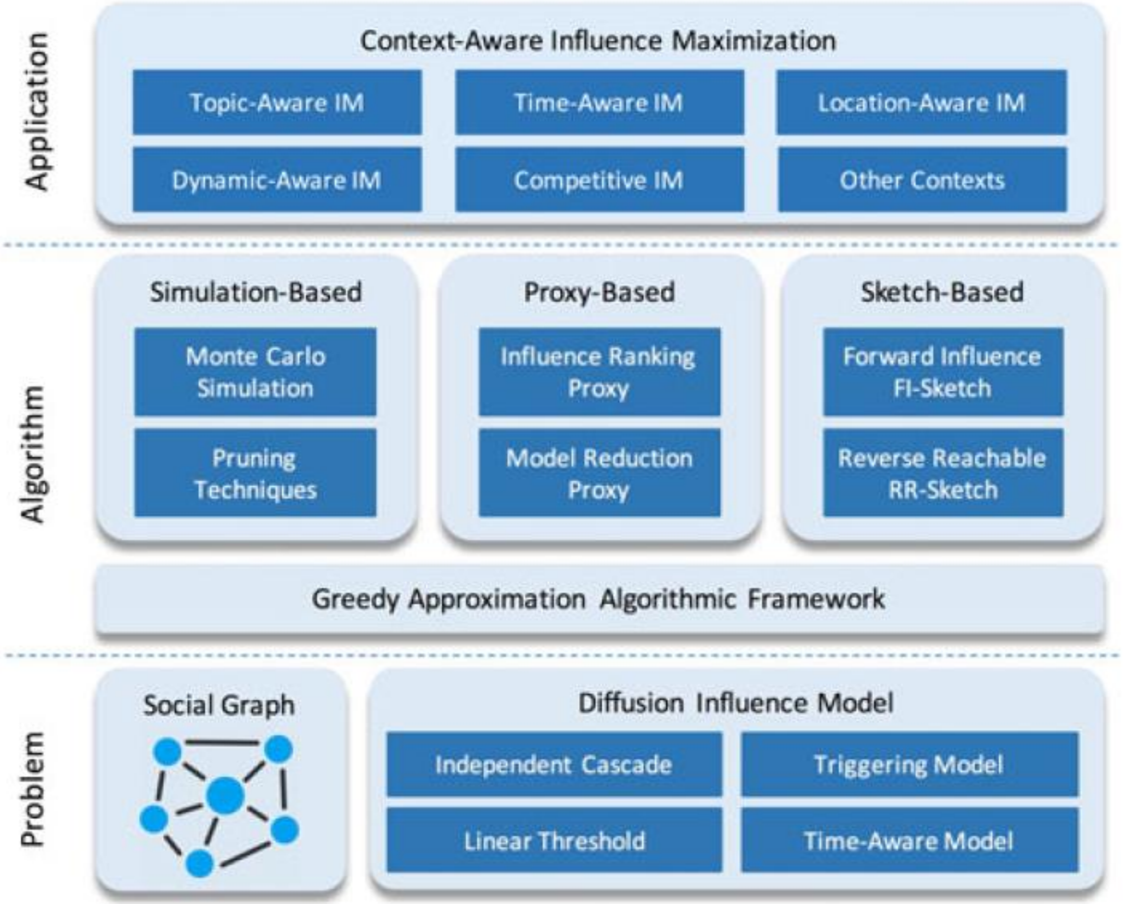
Influence maximization

- Select a set of k users such that the number of influenced users through this set is maximized
- This set of k users is called the seed set
- The expected number of influenced users is the maximum *influence spread*

Challenges in Influence Maximization

- Finding an optimal solution for IM is almost always NP-Hard
- Even evaluating the influence spread of any one seed set is computationally challenging
- Leveraging features of specific social networks, e.g. geolocation

Outline


Application
Context-Aware Influence Maximization
Topic-Aware IM
Time-Aware IM
Location-Aware IM
Dynamic-Aware IM
Competitive IM
Other Contexts
Algorithm
Simulation-Based
Monte Carlo Simulation
Pruning Techniques
Proxy-Based
Influence Ranking Proxy
Model Reduction Proxy
Sketch-Based
Forward Influence FI-Sketch
Reverse Reachable RR-Sketch
Greedy Approximation Algorithmic Framework
Problem
Social Graph
Diffusion Influence Model
Independent Cascade
Triggering Model
Linear Threshold
Time-Aware Model

The Influence Maximization Problem

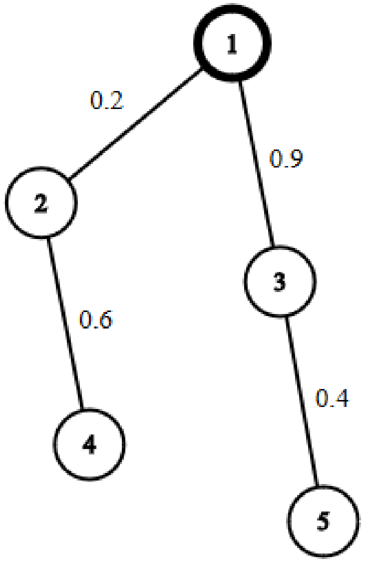
Generic Diffusion Model

- Users are associated with an 'active' or 'inactive' status
- Initially only the seed set is active, everyone else is inactive
- Based on the diffusion model, the set of nodes that S can influence become active
- Then those nodes activate their neighbours, and this continues until all there are no more nodes to activate
- So then $\sigma(S)$, the spread function for S can be defined as the number of nodes with the 'active' status
- We want to find an S s/t $\sigma(S)$, is highest possible value it can be

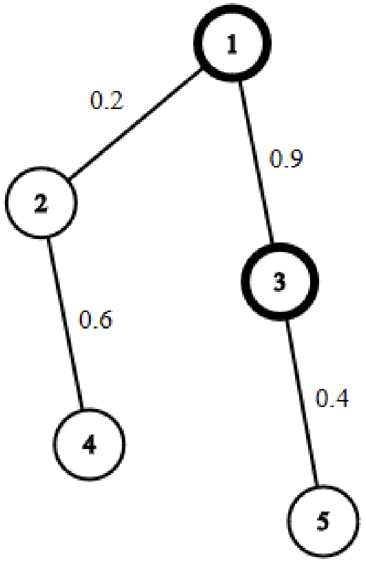
Independent Cascade Model

- Influence probability p(u,v)
- An active edge u will activate its inactive outgoing neighbor v with the probability p(u,v)

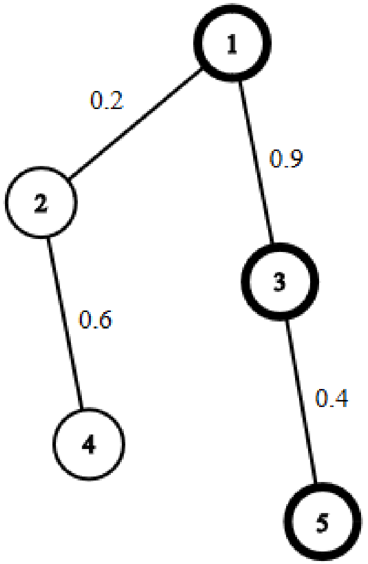
Independent Cascade Model



Independent Cascade Model



Independent Cascade Model



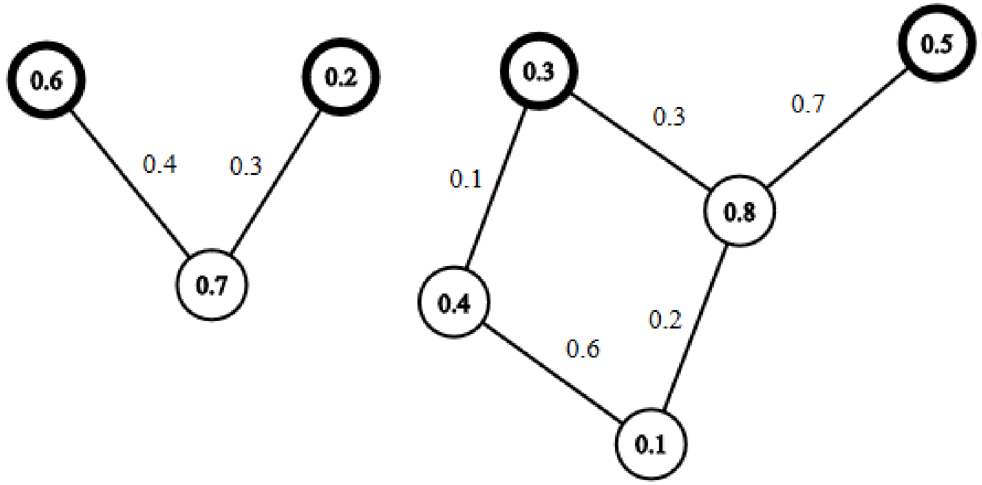
How are the probabilities assigned?

- Earlier → Heuristics e.g. weighted probability assignment
- Now → Learn influence probabilities from data
- Saito et. al. → Expectation maximization to compute the propagation probabilities
- Mathioudakis et. al. → SPINE algorithm to learn the social graph structure and probabilities. Optimal parameters maximize the log likelihood of propagation
- Goyal et. al. → Scalable algorithms
- Kutzkov et. al. → Probabilities in data streams with only one pass

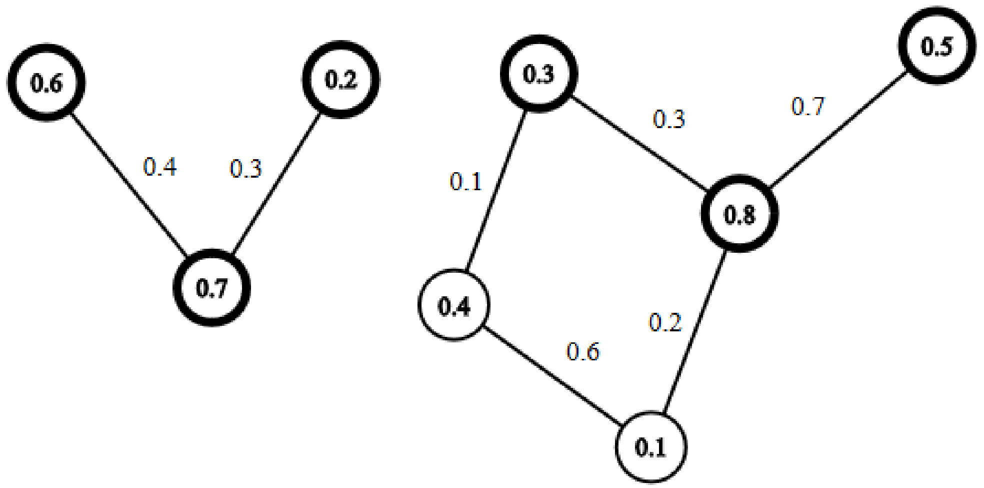
Linear Threshold

- A user can switch its status from inactive to active if a sufficient number of its neighbours are active
- Each edge (u,v) has an associated weight $b_{u,v}$
- Each user v has a threshold θ_v

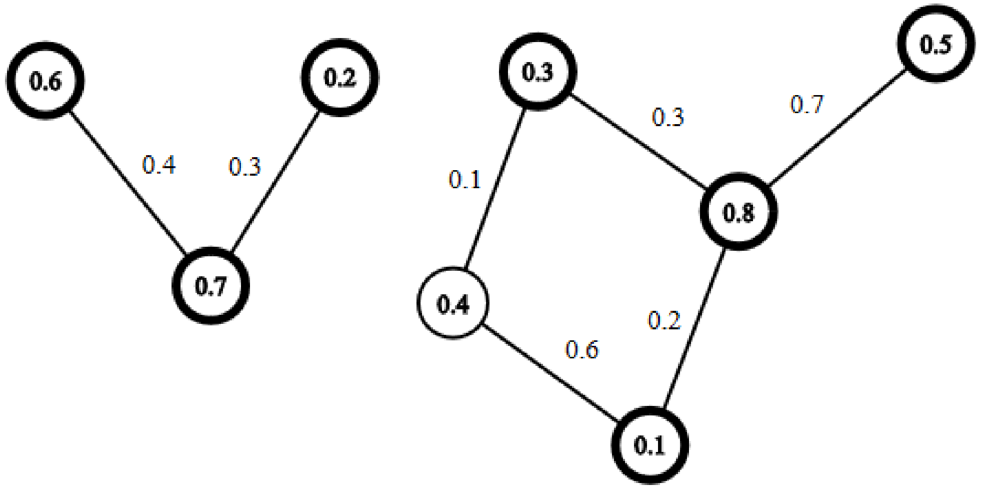
Linear Threshold Model



Linear Threshold Model



Linear Threshold Model



How are the Weights and Thresholds Assigned?

- Weights are assigned using heuristics
- Thresholds are assigned using random sampling
- To the best of the authors' knowledge, there is no data-driven approach for assigning weights

Triggering Model

- Generalised version of IC and LT
- Instead of going by neighbours' weights, a subset of the neighbours is mapped to the probability that that neighbour subset can influence the node. This is the Triggering set.
- Each node has a triggering set. It is chosen randomly
- The initial set S is active at first, everyone else is inactive
- After this, if a node v was inactive in the previous step, and if it has a neighbour in its triggering set that was activated in the previous step, then node v is set to active

Time-Aware Diffusion Models

- Time is a consideration in real-world diffusion models
- Time models are discrete or continuous
- Discrete models → Diffusion happens in discrete steps, like in the previous examples
- Continuous models → considers the pairwise propagation between nodes in a continuous distribution of time

Continuous-Time Model

- Each node u has an activation time t_u
- Each edge (u,v) has a conditional likelihood for activation defined as $p(t_v | t_u; \alpha_{u,v})$, $t_v > t_u$
- $\alpha_{u,v}$ is the parameter of time-aware influence distribution
- Given a stopping time of T > 0, the diffusion instance stops when no more nodes are activated before T
- Exponential model → $p(t_v | t_u; \alpha_{u,v}) = \alpha_{u,v} \cdot e^{-\alpha_{u,v}(t_v - t_u)}$ if $t_v > t_u$ and 0 otherwise

DynaDiffuse Model

- DynaDiffuse model →the propagation rates of nodes decrease exponentially over time
- Consider a node u activated at time t and an edge u → v with a propagation rate of r(u,v)
- The propagation probability for u and v at time t', (t' > t) is $1\text{-}e^{r(u,v).(t'-t)}$

Non-Progressive Models

- In progressive models, once a node is activated, it can’t be set to inactive again
- But this is not the case in non-progressive models
- Not considered in this paper

Problem Hardness of Influence Maximization

- Theorem 1 → The IM problem is NP hard under the IC, LT, TR and CT models
- Proof 1 → To prove NP-Hardness, reduce the problem from a known NP-Hard problem, so then it will be at least as hard as that problem. Under the IC model we reduce from the set cover problem.
- The set cover problem → Given a set of elements U and a collection of m sets S whose union is U, find the smallest number of sets in S whose union is U
- For IM under LT, reduce from the vertex cover problem.
- Can extend these to TR and CT since IC and LT are special cases of TR and CT

Problem Hardness of Influence Maximization

- Theorems 2 and 3 → Computing the influence σ(S) of a seed set S is #P hard under the LT and IC models
- Proof 2 → For LT, reduce from counting simple paths in a directed graph
- Proof 3 → For IC, reduce from counting problem of s – t correctness in a directed graph

Overview of IM Algorithms

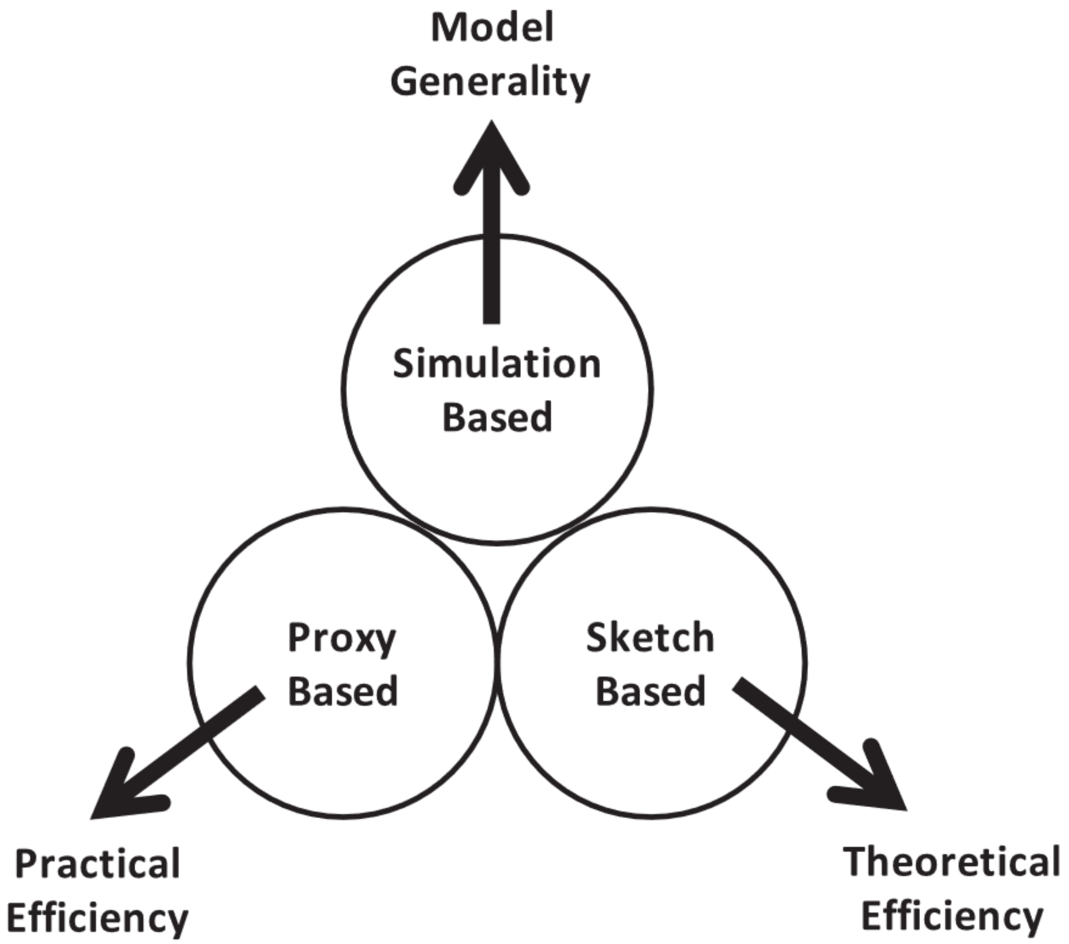
Overview of IM Algorithms

- An optimal solution to the diffusion problem can be approximated if the influence function σ(S) satisfies two properties: monotonicity and submodularity
- Monotonicity → Adding more nodes to a seed set does not reduce its influence spread
- Submodularity → Adding more nodes to a seed set increases its influence spread, but the more the nodes that are added, the smaller these increments become

The Greedy Framework

- Algo initially initialized with a seed set of 0
- Iteratively selects node u into S if node u provides the maximal marginal gain to the influence function $\sigma(S)$. Terminates when there are k distinct nodes in S
- Theorem with the approximation ratio
- Reduce the approximation ratio

Taxonomy of Existing IM Algorithms

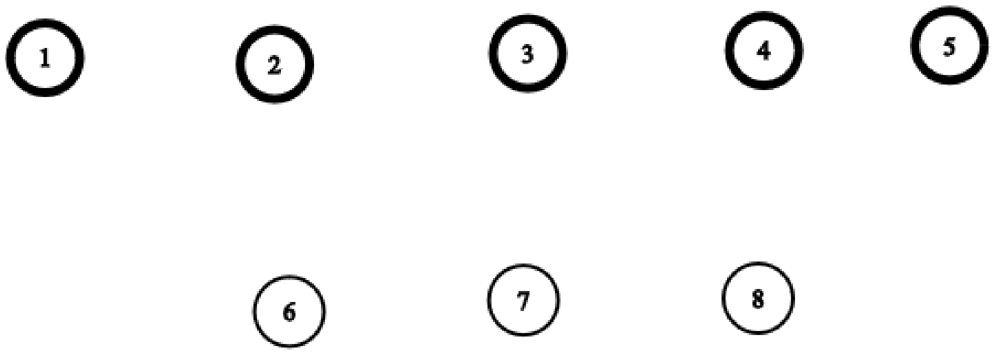


Simulation Algorithms

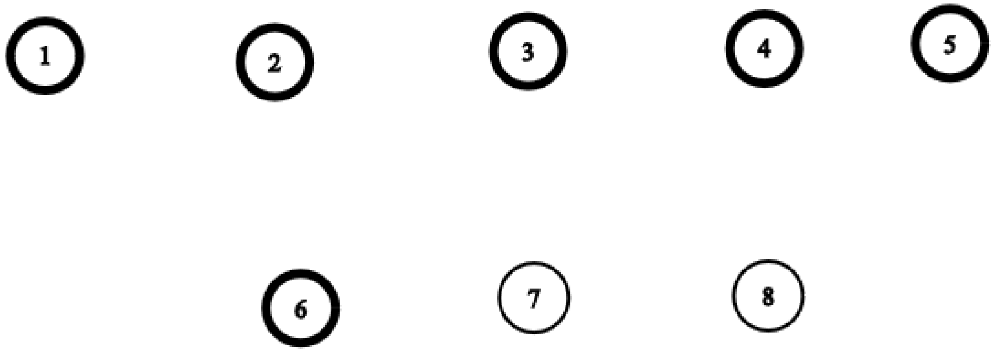
Simulation-Based Approach

- Extension of the greedy framework
- Uses MC simulation to estimate the spread of a seed set
- Takes $\sigma(S_i \cup \{u\})$ → Simulates its activation function wrt the diffusion model → Outputs the number of activated users $I(S_i \cup \{u\})$
- For each $(S_i \cup \{u\})$ the algorithm runs r times and takes the average of $I(S_i \cup \{u\})$ as the estimated spread $\sigma(S_i \cup \{u\})$

Simulation-Based Approach



Simulation-Based Approach



Simulation-Based Approach

- Reduce the number of MC Rounds
  - CELF
  - CELF++
  - UBLF
- Reducing MC Complexity
  - CGA

CELF

- $\Delta(u|S_i) = \sigma(S_i \cup \{u\}) - \sigma(S_i)$ → Marginal gain of u in seed set S_i during iteration i
- According to the supermodularity of the influence function → the marginal influence of u decreases as the seed set increases
- So then, for $Si \subseteq Sj$, $\Delta(u|S_i)$ is an upper bound on $\Delta(u|S_j)$
- Step 1 → Compute the marginal influence for every node wrt an empty seed set and select S_1
- From iteration 2 onwards → Visit each vertex not present in the seed set in descending order of their influence
- Compute the marginal gain wrt the seed set using MC simulations
- If the maximum upper bound of the unvisited nodes is smaller than the maximum influence of the visited nodes → Trigger an early termination and move on to the next iteration

CELF++

- It tweaks the calculation of CELF by considering the influence of a user and the marginal influence of the most influential node before u in the seed set. Isn't much better than CELF

UBLF

- CELF and CELF++ have to calculate the initial marginal influence of each user in the graph which is computationally expensive
- Introduces a value σ' which is a vector that contains the spread σ(u) for each node in the graph
- UBLF derives an upper bound for σ' → $\sigma' \leq \Sigma^{n}_{i=1} PP^{i}$
- PP^{i} is the propagation probability matrix of the graph.
- Since PP is a sparse matrix it, it quickly converges and we can approximate σ'

CGA

- Divide the graph into communities
- The most influential node in each community is added to the seed set

Pros and Cons of Simulation-Based Methods

- Pros
  - Easy to incorporate any diffusion models into these algorithms; it has model generality
  - Strong theoretical property, guaranteed sound results if the solution
- Cons
  - Low computational efficiency

Proxy-Based Approaches

Proxy-Based Approach

- Estimate the influence spread of the seed set instead of running heavy MC simulations
- Make IM algorithms more scalable
- The most practical-oriented methods
- Two types of methods
  - Influence model-based proxy
  - Diffusion model-based proxy

Influence-Ranking Proxy

- Rank all users according to a metric approximating their influences and generate a seed set from the ranking directly
- Simple PageRank Proxy
  - Rank nodes by measures like Pagerank or centrality/betweenness measures; the k highest ones become the seed set
  - Cannot provide fair solutions → influence diverges from degree
  - Does not account for overlaps in influences
- DegDis
  - Ranks users by degree with a discount (DegreeDiscount) for neighbouring nodes in the seed set
  - Only accounts for influence overlaps between neighbours

Influence-Ranking Proxy

- Group PageRank
  - GPR(S) is the sum of the PageRank scores of all the nodes in S with a discount for overlaps and is an upper bound on the estimate of the influence of S
  - Marginal influence of a node can be either calculated through power iterations or be derived from its individual PageRank Score
- IRIE
  - Model the graph of n nodes as a series of n linear equations with n variables, and each variable represents the influence of that node
  - The linear equations are derived as the sum of all influences the node propagates outwards + its influence on itself
  - Solving all n equations each time completes 1 iteration. The equations are updated and solved k times to retrieve the seed set

Influence-Ranking Proxy

- SPIN
  - Models users as players in a coalitional game and captures diffusion as coalition formation
  - The influence of users is evaluated with the Shapely value, a ranking proxy developed for this method
- IMRank
  - Rank the nodes by marginal influence after every iteration using a Last-To-First strategy until the ranking converges
- Problems with Influence Ranking Proxies
  - The metrics used here may not always be a good fit for the IM problem and influence estimation may seriously diverge
  - The properties of diffusion are completely ignored here
  - These problems are addressed with diffusion model proxies

Diffusion Model Proxies

- Simplify the diffusion process to address the #P-hardness of evaluating the influence function
- Can be done by:
  - Reducing stochastic diffusion models (IC, LT) to a deterministic model where the influence spread of a seed set can be computed exactly
  - Restrict the influence of a user to a small local subgraph containing u and ignoring the rest
- Once the proxy model is computed, the algorithm from the greedy framework is employed

IC Diffusion Model Reduction Proxy

- Focus is on reducing the complexity of influence estimation by considering only the significant probabilistic paths
- SPM and SP1M
  - Consider the shortest path from a node to any node in the seed set, i.e. d(S,v)
  - In SPM, v can only be activated in step d(S,v)
  - In SP1M, v can be activated in d(S,v) or d(S,v) + 1
  - Simple computations but performance is poor if path probabilities are not constant or not small

IC Diffusion Model Reduction Proxy

- MIA (Maximum influence arborescence)
  - Restrict influence of a node to a local tree structure, the root of which is that node
  - For a pair of nodes (u,v), u can only influence v through a maximum threshold path greater than a selected threshold
  - Use Djikstra's shortest path algorithm to compute the in-arboresence and out-arborescence for a user given a threshold value. The influence of a node is calculated with these values
- PMIA
  - A problem with MIA is that nodes can block the influence of other nodes. E.g. if we're considering the shortest path (a→b) and c is in the middle, c blocks a's influence.
  - PMIA updates the influenced in-arborescence after adding a node to the seed set to prevent this
- Both MIA and PMIA are not scalable and do not perform well on large, dense graphs

LT Diffusion Model Reduction Proxy

- LDAG
  - Restrict the graph to a DAG and compute the influence more effectively
  - The problem of finding the optimal DAG itself is NP Hard
  - LDAG introduces a loss in quality
- SIMPATH
  - The influence of a node is calculated by summing up all its paths with the probability greater than a threshold value
  - Optimizes execution by finding the vertex cover set at the beginning and calculating the influence of those nodes and in later rounds starts with calculating the influence of promising nodes first
  - For larger graphs, and seed sets, LDAG could outperform SIMPATH
- EASYSIM
  - Estimates influence of a node by counting influence paths within length l + accounts for overlaps between different paths

Pros and Cons of Proxy-Based Modelling

- Pros
  - More rooted in practical applications and with a focus on efficiency
- Cons
  - Lack of theoretical guarantee
  - These methods are generally not stable when it comes to unstable scenarios

Sketch-Based Algorithms

Sketch-Based Algorithms

- Improve theoretical efficiency while maintaining the approximation guarantee
- Problem with simulation → Runs MC many times on many nodes
- Avoid this by computing sketches based on the diffusion model

Comparison of Sketch techniques

FORWARD INFLUENCE SKETCH

- Construct a subgraph by running an influence process w.r.t. a diffusion model. Estimate the influence of a set S by constructing multiple sketches and taking the average of the users reached in each sketch
- Needs to visit all the nodes
- Prohibitively high worst-case complexities

REVERSE REACHABLE SKETCH

- A Reverse Reachable set is constructed by generating random nodes and including the ones that can be reached by a specific node, in our case, nodes in the seed set.
- Construct multiple RR sketches for multiple random nodes and add them to a set called the Random RR set.
- Influential nodes are more likely to show up in these sketches, and these can form the seed set

Forward Influence Sketches

- NewGreIC → At each iteration of the greedy framework, constructs a number of sketches evaluating the gain of each node not in the seed set simultaneously
- StaticGreedy → Constructs θ sketches to perform all influence evaluations
- StaticGreedyDU → At the end of an iteration, after a seed set is obtained, prune all the users reached by that seed set
- PrunedMC → For each sketch, build a DAG in which each node is an SCC. Build an index structure for ancestors and descendants. Makes it easier to evaluate the influence of a node if it is the ancestor of a hub
- SKIM → Perform BFS walks on sketches constructed by using bottom- K^2 minHash values to update these values for multiple candidate sets at once i.e. evaluating the spread of that node

Reverse Reachable Sketch

- RIS → Keep generating random RR sets until the total number of edges examined during the generation process reaches a pre-defined threshold value
- TIM → Improves the complexity of RIS by introducing a parameter by introducing OPT, the parameter for the optimal seed set. TIM+ has the same worst case complexity but better empirical results. IMM uses a better technique for estimating OPT and is more efficient than TIM/TIM+

Reverse Reachable Sketch

- BKRIS → Uses a lazy sampling technique by first estimating a lower bound on OPT and then deriving a sample size of sketches from it. It uses RR and avoids materializing all the sketches unless needed. It is easier on the memory than the previous methods discussed
- SSA → An orthogonal stop-and stare optimization that improves on IMM. Whenever the influence of a seed step is close to that of the previous iteration, execution stops and the seed set is returned.
- Tang et. al. → Use Djikstra's shortest path algorithm with the influence distribution from the CT model to extend the methods previously discussed to CT models

Pros and Cons of Sketch-Based Algorithms

- Pros
  - Theoretical guarantee with proven low complexities
- Cons
  - The constructed sketches must align with the underlying diffusion model, meaning that results from this approach may not be widely applicable

Context-Aware Influence Maximization

Topic-Aware Influence Maximization

- Topics i.e. item characteristics or user preferences are taken into consideration when determining the influence of a seed set
- IM for Topic-Relevant Targets
  - Maximize the influence over users who are relevant to the query topics i.e. the topic-relevant targets
  - The concept of benefit is used to differentiate users
  - The summation of benefits of the activated users becomes the influence spread (the targeted influence)
  - The aim becomes to find techniques to maximize the targeted influence
  - Li et. al., Nguyen et. al., personalized IM problem

Topic-Aware Influence Maximization

- IM for Topic-Dependent Diffusion
  - The edges between every pair of nodes is topic-dependent. E.g. a v might be activated by u in sports but not in politics
  - Modifies the IC model by introducing topic-specific edge probabilities
  - The challenge here is the enormous number of potential queries and the results that lead to different probabilistic graphs
  - INFLEX → Two queries that have similar topic distributions will also have similar influence spread
  - Chen et. al. → Estimate an upper bound of the influence for each user and only compute the influence for users with higher upper bounds. Precompute seeds for offline-sampled topic distributions

Time-Aware Influence Maximization

- Diffusion instances may not stop at a particular fixed time, especially in continuous diffusion models
- Time-Aware IM aims put a time constraint on the diffusion process
- IC-M → Finds the optimal seed set over random processes of at most Tau steps
- LAIC/CT-IC → Two similar models developed concurrently. They introduce the concept of a time delay distribution in conjunction with propagation probability
- Continuous-Time Independent Cascade Model → Uses the greedy framework with lazy forward optimization under a CT diffusion model

Location-Aware Influence Maximization

- Maximize the influence of location-relevant users instead of any generic users
- Solutions tend to combine IM algorithms with spatial index schemes
- Li et al. → Uses a best-first search framework that preferentially accesses users with large upper bounds. The pruning is done with the help of a QuadTree spatial index
- Wang et. al. → Maximize influence spread weighted by a users' distance to an query location. Anchor locations are judiciously selected and used in bound estimation for queries

Dynamic Influence Maximization

- Incrementally process changes of the social graph
- Zhuang et. al. → (Assumption) Changes in graphs can only be detected periodically by probing a small number of nodes. The proposed solution of probing nodes, generating subgraphs and finding the seed set from the subgraph does not align with IC/LT
- UBI → Models dynamics of the network as a sequence of snapshots and updates the seed set it obtains from UBLF and SPIM from the most recent snapshot
- Ohsaka et. al. → Adds and deletes nodes (EXPANDS or SHRINKS) by resampling based on changes it receives about the graph in real time

Competitive Influence Maximization

- IM for scenarios in which one's own influenced is increased while opponents' influence is minimized
- Known Opponent Strategies → Influence Blocking Maximization (E.g. countering misinformation)
  - Zhu et. al. → A modified version of IC in which a node can serve as the seed for multiple diffusions
- Unknown Opponent Strategies (More realistic scenario)
  - Li et. al → Model the problem as a multiparty game and reach the objective of finding a Nash equilibrium strategy for each player that maximizes their influence
- Comparative IM (NP Hard)
  - Competition → If a node is influenced by A it is less likely to be influenced by B (SELFINFMAX)
  - Complimentary → If a node is influenced by A it is more likely to be influenced by B (COMPINFMAX)

Research Challenges and Directions

Research Challenges and Directions

- Determining stability of IM algorithms
  - IM algorithms have poor stability in the face of noisy input probabilities
  - Results are highly sensitive to this noise
- Breaking the boundary of submodularity
  - The submodularity requirement for any IM solution is too restrictive in some cases, e.g. Opinion Aware IM
  - With non-submodularity, the greedy framework does not apply, meaning more general functions are needed
  - Weakly submodular functions are better for modelling real-world applications
- Considering the Group Norm
  - Study how group dynamics and norms affect peoples' behaviour and what they are influenced by

Conclusion

Conclusion

- The concepts of IM were discussed
- The taxonomy for IM and IM-related works were presented
- Points for potential future research were discussed

Citation Analysis

This Paper Cited…

- A. Arora, S. Galhotra, and S. Ranu, “Debunking the myths ofinfluence maximization: An in-depth benchmarking study,”
  - It is an interesting benchmarking survey that shows that purported improvements to algorithms are not actually much better (E.g. CELF++ to CELF). It also concludes that there is no single ‘best’ IM algorithm
- W. Lu, X. Xiao, A. Goyal, K. Huang, and L. V. S. Lakshmanan, “Refutations on ”debunking the myths of influence maximization: An in-depth benchmarking study”,”
  - Disproves almost everything stated in the previous paper,

This Paper Cited…

- Goldenberg, B. Libai, and E. Muller, “Talk of the network: A complex systems look at the underlying process of word-of-mouth,”
  - A non-graph take on word of mouth, it uses agent-based modelling to simulate a network and discusses IM’s applications from a marketing standpoint
- Y. Li, J. Fan, D. Zhang, and K. Tan, “Discovering your sellingpoints: Personalized social influential tags exploration,”
  - This paper applies the concepts of IM in a unique way: for any user of a social media site, its algorithm produces the set of tags best suited to boosting your online profile
- H. C. Kelman, “Compliance, identification, and internalization three processes of attitude change,”J. Conflict Resolution, vol. 2,no. 1, pp. 51–60, 1958
  - A highly influencial communications paper published in 1958 in the Journal of Conflict Resolution that still has relevance today because of the way it explains conformity as a part of group dynamics.

This Paper has been Cited by…

- Zhang, Ping, Zhifeng Bao, Yuchen Li, Guoliang Li, Yipeng Zhang, and Zhiyong Peng. "Trajectory-driven influential billboard placement."
  - This proposes the problem: given a limited budget and a limited number of billboards, which are the best spots for the billboards to maximise their impact?
- Iyer, Shankar, and Lada A. Adamic. "The Costs of Overambitious Seeding of Social Products."
  - This paper argues that over-seeding to maximize influence does not necessarily guarantee good results, and that gradual influence increasing techniques may work better
- Chen, Huiping, Grigorios Loukides, Jiashi Fan, and Hau Chan. "Limiting the Influence to Vulnerable Users in Social Networks: A Ratio Perspective."
  - This paper discusses a drawback of IM which is that it can influence users towards something that might cause them harm and propose an algorithm to protect vulnerable users from such influences

This Paper has been Cited by…

- Tsang, Alan, Bryan Wilder, Eric Rice, Milind Tambe, and Yair Zick. "Group-fairness in influence maximization." *arXiv preprint arXiv:1903.00967* (2019).
  - This paper solely focuses on fairness and ethics in influence maximization and using legal and game-theory terminology proposes an algorithmic framework for determining if an IM technique is fair or not
- Banerjee, Suman, Mamata Jenamani, Dilip Kumar Pratihar, and Abhinav Sirohi. "A priority-based ranking approach for maximizing the earned benefit in an incentivized social network." In *International Conference on Intelligent Systems Design and Applications*, pp. 717-726. Springer, Cham, 2018.
  - A practical variant of the IM problem in which selecting nodes into the seed set has a cost, and there is a budget to limit the number of nodes in the seed set. In addition to maximizing spread, the problem also becomes how to optimize the budget.