Information and Influence Propagation in Social Networks: Modeling and Influence Maximization

Wei Chen 陈卫
Microsoft Research Asia
Social influence and viral phenomena
Voting mobilization: A Facebook study

- Voting mobilization [Bond et al, Nature’2012]
  - show a Facebook msg. on voting day with faces of friends who voted
  - generate 340K additional votes due to this message, among 60M people tested
Influence Propagation Modeling and Influence maximization task

- Studies the stochastic models on how influence propagates in social networks
  - Its properties, e.g. submodularity
- Influence maximization: given a budget $k$, select at most $k$ nodes in a social network as seeds to maximize the influence spread of the seeds
  - Applications in viral marketing, diffusion monitoring, rumor control, etc.
Outline of This Talk

• Basic concepts: influence diffusion models, influence maximization task, submodularity, greedy algorithm
• Scalable algorithm based on reverse influence sampling (RIS)
• Influence-based centrality measures
  – Shapley centrality
  – Single Node Influence (SNI) centrality
• Other models and tasks
Independent cascade model

- Directed graph $G = (V, E)$
- Each edge $(u, v)$ has an influence probability $p(u, v)$
- Initially seed nodes in $S_0$ are activated
- At each step $t$, each node $u$ activated at step $t - 1$ activates its neighbor $v$ independently with probability $p(u, v)$
- Influence spread $\sigma(S)$: expected number of activated nodes
- Correspond to bond percolation
Linear threshold model

• Each edge \((u, v)\) has an influence weight \(w(u, v)\):
  – when \((u, v) \notin E, w(u, v) = 0\)
  – \(\sum_u w(u, v) \leq 1\)

• Each node \(v\) selects a threshold \(\theta_v \in [0,1]\) uniformly at random

• Initially seed nodes in \(S_0\) are activated

• At each step, node \(v\) checks if the weighted sum of its active in-neighbors is greater than or equal to its threshold \(\theta_v\), if so \(v\) is activated
Interpretation of IC and LT models

• IC model reflects simple contagion, e.g. information, virus

• LT model reflects complex contagion, e.g. product adoption, innovations (activation needs social affirmation from multiple sources [Centola and Macy, AJS 2007])

• More general models are studied: triggering model, general threshold models, decreasing cascade model, etc.
  – Note: not all models correspond to reachability on random graphs, e.g. general threshold model corresponds to random hyper-graphs (ongoing research)
Influence maximization

- Given a social network, a diffusion model with given parameters, and a number $k$, find a seed set $S$ of at most $k$ nodes such that the influence spread of $S$ is maximized.
- NP-hard
- Based on submodular function maximization
- [Kempe, Kleinberg, and Tardos, KDD’2003]
Submodular set functions

- **Sumodularity** of set functions \( f : 2^V \rightarrow R \)
  - for all \( S \subseteq T \subseteq V \), all \( v \in V \setminus T \),
    \[
    f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)
    \]
  - diminishing marginal return
  - an equivalent form: for all \( S, T \subseteq V \)
    \[
    f(S \cup T) + f(S \cap T) \leq f(S) + f(T)
    \]
- **Monotonicity** of set functions \( f \): for all \( S \subseteq T \subseteq V \),
  \[
  f(S) \leq f(T)
  \]
- Influence spread function \( \sigma(S) \) is monotone and submodular in the IC model (and many other models)
Example of a submodular function and its maximization problem

• set coverage
  – each entry $u$ is a subset of some base elements
  – coverage $f(S) = | \bigcup_{u \in S} u |$
  – $f(S \cup \{v\}) - f(S)$: additional coverage of $v$ on top of $S$

• $k$-max cover problem
  – find $k$ subsets that maximizes their total coverage
  – NP-hard
  – special case of IM problem in IC model

sets

elements
Submodularity of influence diffusion models

- Based on equivalent live-edge graphs

**Diffusion Dynamic**

\[ \Pr(\text{set A is activated given seed set S}) \]

**Random Live-Edge Graph: Edges are randomly removed**

\[ \Pr(\text{set A is reachable from S in random live-edge graph}) \]
Random live-edge graph for the IC model and its reachable node set

• Random live-edge graph in the IC model
  – each edge is independently selected as live with its influence probability
• Pink node set is the active node set reachable from the seed set in a random live-edge graph
• Equivalence is straightforward (it is essentially bond percolation)
Random live-edge graph for the LT model and its reachable node set

- Random live-edge graph in the LT model
  - each node select at most one incoming edge, with probability equal to its influence weight
- Pink node set is the active node set reachable from the seed set in a random live-edge graph
- Equivalence is based on uniform threshold selection from [0,1], and linear weight addition
- Not exactly a bond percolation
Submodularity of influence diffusion models (cont’d)

• Submodularity of $|R(\cdot, G_L)|$
  • for any $S \subseteq T \subseteq V$, $v \in V \setminus T$,
  • if $u$ is reachable from $v$ but not from $T$, then
  • $u$ is reachable from $v$ but not from $S$
  • Hence, $|R(\cdot, G_L)|$ is submodular
• Therefore, influence spread $\sigma(S)$ is submodular in the IC model
Greedy algorithm for submodular function maximization

1: initialize $S = \emptyset$
2: for $i = 1$ to $k$ do
3: select $u = \arg \max_{w \in V \setminus S} [f(S \cup \{w\}) - f(S)]$
4: $S = S \cup \{u\}$
5: end for
6: output $S$
Property of the greedy algorithm

• Theorem: If the set function $f$ is monotone and submodular with $f(\emptyset) = 0$, then the greedy algorithm achieves $(1 - 1/e)$ approximation ratio, that is, the solution $S$ found by the greedy algorithm satisfies:

$$f(S) \geq \left(1 - \frac{1}{e}\right) \max_{S' \subseteq V, |S'| = k} f(S')$$
Hardness of Influence Maximization and Influence Computation

• In IC and LT models, influence maximization is NP-hard
  – IC model: reduction from the set cover problem
• In IC and LT models, computing influence spread $\sigma(S)$ for any given $S$ is #P-hard [Chen et al. KDD’2010, ICDM’2010].
  – IC model: reduction from the s-t connectedness counting problem.
• Implication of #P-hardness of computing $\sigma(S)$
  – Greedy algorithm needs adaptation --- using Monte Carlo simulations
MC-Greedy: Estimating influence spread via Monte Carlo simulations

• For any given $S$
• Simulate the diffusion process from $S$ for $R$ times ($R$ should be large)
• Use the average of the number of active nodes in $R$ simulations as the estimate of $\sigma(S)$
• Can estimate $\sigma(S)$ to arbitrary accuracy, but require large $R$
  – Theoretical bound can be obtained using Chernoff bound.
Theorems on MC-Greedy algorithm

**Theorem 3.6** Let $S^* = \arg\max_{|S| \leq k} f(S)$ be the set maximizing $f(S)$ among all sets with size at most $k$, where $f$ is monotone and submodular, and $f(\emptyset) = 0$. For any $\varepsilon > 0$, for any $\gamma$ with $0 < \gamma \leq \frac{\varepsilon/k}{2 + \varepsilon/k}$, for any set function estimate $\hat{f}$ that is a multiplicative $\gamma$-error estimate of set function $f$, the output $S^g$ of $\text{Greedy}(k, \hat{f})$ guarantees

$$f(S^g) \geq \left(1 - \frac{1}{e} - \varepsilon\right) f(S^*).$$

**Theorem 3.7** With probability $1 - 1/n$, algorithm MC-Greedy$(G, k)$ achieves $(1 - 1/e - \varepsilon)$ approximation ratio in time $O(\varepsilon^{-2} k^3 n^2 m \log n)$, for both IC and LT models.

- Polynomial time, but could be very slow: 70+ hours on a 15k node graph
Simulation on Real Network NetHEPT

- NetHEPT: collaboration network on arxiv
- MC-Greedy[20000] is the best
- MC-Greedy[200] is worse than Degree
- Random is the worst

uniform IC: $p=0.01$

weighted IC: $p(u, v) = 1/d_v^{in}$

- Number of nodes: 15233
- Number of edges with duplicated edges: 58891
- Number of edges: 31398
- Average degree: 4.12
- Maximal degree: 64
- Number of connected components: 1781
- Largest component size: 6794
- Average component size: 8.55
### Probabilists’ View vs. Computer Scientists’ View on Diffusion

<table>
<thead>
<tr>
<th></th>
<th>Probabilists’ view</th>
<th>Computer scientists’ view</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>subject</strong></td>
<td>(stochastic) diffusion on random networks</td>
<td>(stochastic) diffusion on fixed networks (often equivalent to deterministic diffusion on random sub-networks of the fixed network)</td>
</tr>
<tr>
<td><strong>network</strong></td>
<td>family of random networks ($n \to \infty$, e.g. configuration model), infinite lattice, etc.</td>
<td>fixed network with arbitrary topology</td>
</tr>
<tr>
<td><strong>diffusion models</strong></td>
<td>percolation, SIR, SIS, etc.</td>
<td>independent cascade (equivalent to bond percolation), linear threshold, triggering, general threshold, etc.</td>
</tr>
<tr>
<td><strong>goal</strong></td>
<td>reveal properties of the diffusion, e.g. condition of the phase transition</td>
<td>optimization, e.g. influence maximization</td>
</tr>
<tr>
<td><strong>method and tools</strong></td>
<td>probabilistic analysis, Markov process, branching process,</td>
<td>submodularity analysis, submodular maximization, concentration inequalities</td>
</tr>
<tr>
<td><strong>focus</strong></td>
<td>probabilistic analysis, phase transition condition, size distribution, etc.</td>
<td>algorithm design, efficiency, approximation ratio</td>
</tr>
</tbody>
</table>
Outline of This Talk

• Basic concepts: influence diffusion models, influence maximization task, submodularity, greedy algorithm
• Scalable algorithm based on reverse influence sampling (RIS)
• Influence-based centrality measures
  – Shapley centrality
  – Single Node Influence (SNI) centrality
• Other models and tasks
Ways to improve scalability

- Fast deterministic heuristics
  - Utilize model characteristic
  - MIA/IRIE heuristic for IC model [Chen et al. KDD’10, Jung et al. ICDM’12]
  - LDAG/SimPath heuristics for LT model [Chen et al. ICDM’10, Goyal et al. ICDM’11]

- Monte Carlo simulation based
  - Lazy evaluation [Leskovec et al. KDD’2007], Reduce the number of influence spread evaluations

- New approach based on Reverse Influence Sampling (RIS)
  - First proposed by Borgs et al. SODA’2014
  - Improved by Tang et al. SIGMOD’14, 15 (TIM/TIM+, IMM), Nguyen et al. SIGMOD’16 (SSA/D-SSA), Nguyen et al. ICDM’17 (SKIS), Tang et al. SIGMOD’18 (OPIM)
Key Idea: Reverse Influence Sampling

- Reverse Reachable sets: (use IC model as an example)
  - Select a node \( v \) uniformly at random, call it a root
  - From \( v \), simulate diffusion, but in reverse order --- every edge direction is reversed, with same probability
  - The set of all nodes reached (including \( v \)) is the reverse reachable set \( R \) (rooted at \( v \)).

- Intuition:
  - If a node \( u \) often appears in RR sets, it means that if using \( u \) as the seed, its influence is large --- efficiently collect evidence of influencers

- Technical guarantee: For any seed set \( S \),
  \[ \sigma(S) = n \cdot Pr\{S \cap R\} \]

- [Borgs et al. SODA'2014]
• Collect all RR sets
• Greedily find top $k$ nodes cover most number of RR sets
How to Decide the Number of RR Sets:
IMM: Influence Maximization via Martingales

• Estimate a lower bound on the optimal influence spread
  – Repeated halving the estimate, double the RR sets
  – Use greedy on RR sets to get a lower bound solution
  – Verify if it is close to the estimate
  – Generate final number of RR sets

• Use greedy on the RR sets to find $k$ nodes that cover the most number of RR sets
IMM Theoretical Result

- Theorem: For any $\epsilon > 0$ and $\ell > 0$, IMM achieves $1 - \frac{1}{e} - \epsilon$ approximation of influence maximization with at least probability $1 - \frac{1}{n^\ell}$. The expected running time of IMM is $O \left( \frac{(k+\ell)(m+n)\log n}{\epsilon^2} \right)$.

- Martingale based probabilistic analysis
  - RR sets are not independent --- early RR sets determine whether later RR sets are generated --- form a Martingale

Near linear time to graph size
IMM Empirical Result

- LiveJournal: blog network
  - $n = 4.8M$
  - $m = 69.0M$
- Orkut: social network
  - $n = 3.1M$
  - $m = 117.2M$
- $\varepsilon = 0.5$, $\ell = 1$
- IC model, $p(u, v) = 1/d_{v}^{\text{in}}$
  - $d_{v}^{\text{in}}$: indegree of $v$
RIS Summary

• **Advantages**
  – Theoretical guarantee
  – RIS approach can be applied to many other situations
  – Easily tuned between theoretical guarantee and practical efficiency (by tuning $\varepsilon$)

• **Issues**
  – Memory bottleneck (need to store all RR sets)

• **Different RIS-based algorithm improve on different ways of estimating the number of RR sets needed**
Scalable Influence Maximization Trilemma

- Quality guarantee
- Time efficiency
- Memory efficiency

Monte Carlo greedy algorithms
RIS-based algorithms
Graph heuristics
Outline of This Talk

• Basic concepts: influence diffusion models, influence maximization task, submodularity, greedy algorithm
• Scalable algorithm based on reverse influence sampling (RIS)
• Influence-based centrality measures
  – Shapley centrality
  – Single Node Influence (SNI) centrality
• Other models and tasks
Influence-based Centrality Measures

• Network centrality is a key concept in network science
• Most existing network centrality is structure-based: degree centrality, closeness centrality, betweenness centrality, etc.
• When we care about influence propagation in the network, we should look into influence-based centrality
  – [Chen and Teng, WWW’2017]
  – Define two influence-based centrality: Shapley centrality and Single-Node-Influence centrality
  – Provide an axiomatic study on the two centrality measures
  – Provide a scalable algorithmic framework for computing the two centralities
Cooperative Game Theory and Shapley Value

- Measure individual power in group settings

- Cooperative game over $V = [n]$, with characteristic function $\tau: 2^V \rightarrow \mathbb{R}$
  - $\tau(S)$: cooperative utility of set $S$

- Shapley value $\phi: \{\tau\} \rightarrow \mathbb{R}^n$:
  \[
  \phi_v(\tau) = \mathbb{E}_\pi[\tau(S_{\pi,v} \cup \{v\}) - \tau(S_{\pi,v})] = \frac{1}{n!} \sum_{\pi \in \Pi} (\tau(S_{\pi,v} \cup \{v\}) - \tau(S_{\pi,v}))
  \]
  - $\Pi$: set of permutations of $V$
  - $S_{\pi,v}$: subset of $V$ ordered before $v$ in permutation $\pi$
  - Average marginal utility on a random order

- Enjoy a unique axiomatic characterization
Shapley Centrality

• Node $v$'s Shapley Centrality is the Shapley value of the influence spread function

$$\psi_v^{Shapley}(\mathcal{I}) = \phi_v(\sigma \mathcal{I})$$

– Treat influence spread function as a cooperative utility function

• Measure node’s irreplaceable power in groups
• More precisely, node’s marginal influence in a random order
• Shapley centrality can be uniquely characterized by five axioms (omitted)
• Scalable algorithm for Shapley centrality computation exists, based on RIS approach
Key Observation Linking RR Sets with Shapley Value

- Let \( R \) be a random RR set
  \[
  \psi_u^{\text{Shapley}} = n \cdot \mathbb{E}_R[\mathbb{I}\{u \in R\}/|R|]
  \]
- If \( u \) is not in \( R \) rooted at \( v \), \( u \) has no marginal influence
- If \( u \) is in \( R \) root at \( v \),
  - If \( u \) is ordered after any other node in \( R \) in a random permutation, \( u \) has no marginal influence to \( v \)
  - If \( u \) is ordered before all other nodes in \( R \) in a random permutation, \( u \) has marginal influence of 1 to \( v \); this happens with probability \( 1/|R| \)
  - \( v \) is uniformly chosen, so total marginal influence multiplied by \( n \)
Scalable Algorithm for Shapley Centrality

• Use a similar algorithmic structure as IMM
• Same algorithmic structure can be used to compute other influence-based centralities, such as Single-Node-Influence centrality, propagation-distance based centrality [Chen, Teng and Zhang, 2018], etc.

• A big advantage over RIS-based influence maximization algorithms:
  – No memory overhead --- no need to store RR sets:
    • Generate one RR set $R$, for each node $u \in R$, cumulate its score with $1/|R|$
Outline of This Talk

• Basic concepts: influence diffusion models, influence maximization task, submodularity, greedy algorithm
• Scalable algorithm based on reverse influence sampling (RIS)
• Influence-based centrality measures
  – Shapley centrality
  – Single Node Influence (SNI) centrality
• Other models and tasks
Example 1: Influence Propagation with Negative Opinions

- Quality factor $q$
  - If a node is positively influenced, with probability $q$ it turns positive and probability $1 - q$ it turns negative.
  - Both positive and negative influence propagates as in the IC model.
  - Negative influence only activates nodes in the negative state.
- Model negative opinion due to quality defect.
  - Model negativity bias: people are more likely to believe negative opinions than positive opinions.
- Satisfy submodularity, could be made scalable.
- [Chen et al. SDM’2011]
Example 2: Influence Blocking Maximization

• Two competitive items A and B
  – A wants to block the propagation of B as much as possible
  – Application: rumor control

• Competitive diffusion model
  – Competitive IC model: may not be submodular
  – Competitive LT model: submodular

• [Budak et al. WWW’2011, He et al. SDM’2012]
Example 3: Complementary Diffusion Model

- Two items A and B, with global adoption parameters (GAP)
  - $q_{A|\emptyset}$: probability of adopting A when not adopted anything yet
  - $q_{B|\emptyset}$: probability of adopting B when not adopted anything yet
  - $q_{A|B}$: probability of adopting A when B is already adopted
  - $q_{B|A}$: probability of adopting B when A is already adopted
  - $q_{A|\emptyset} \geq q_{A|B}$, $q_{B|\emptyset} \geq q_{B|\emptyset}$: mutually competitive
  - $q_{A|\emptyset} \leq q_{A|B}$, $q_{B|\emptyset} \leq q_{B|\emptyset}$: mutually complementary

- Diffusion follows the IC model
- Self-maximization and complementary-maximization
- Boundary cases are submodular, other cases are not submodular
  - Apply sandwich optimization for non-submodular cases
- [Lu et al. SIGMOD’2016, Zhang and Chen, TCS’2018]
Conclusion and Future Work

• Influence maximization has rich internal problems and external connections to study
  – many optimization, learning and game theoretic studies can be instantiated on the influence maximization task

• Many possible new directions, beyond summarized already
  – Non-submodular influence maximization (e.g. [Zhang et al. KDD’14, Chen et al. EC’15, Lu et al. SIGMOD’16, Lin et al. ICDE’17, Li et al. NIPS’18])
  – Influence maximization in dynamic networks

• Influence maximization with phase transition / percolation?

• Need validations on large-scale real social networks
Reference Resources

- Search “Wei Chen Microsoft”
- KDD’12 tutorial on influence spread in social networks
- my papers and talk slides
- A recent survey on influence maximization [Li et al. TKDE’2018]
Thanks!