Viral Marketing in Social Networks

Dr. Ramasuri Narayanam
Senior Research Scientist
Adobe Research - India
Email: rnarayanam@adobe.com

19-December-2023

Adobe

Outline of the Presentation

1. **Viral Marketing with Single Product**
2. Viral Marketing - Extensions
3. Viral Marketing with Product Cross-sell
4. Summary

Key Resources of this Topic

- D. Kempe, J.M. Kleinberg, and E. Tardos. *Maximizing the spread of influence through a social network*. In SIGKDD, 2003.
- Ramasuri Narayanam and Y. Narahari. *A Shapley Value based Approach to Discover Influential Nodes in Social Networks*. In IEEE Transactions on Automation Science and Engineering (IEEE TASE), 2011.
- Ramasuri Narayanam and Y. Narahari. *Determining Top-k Nodes in Social Networks using the Shapley Value*. In AAMAS, pages 1509-1512, Portugal, 2008.
- Ramasuri Narayanam and Amit A. Nanavati. *Viral Marketing for Product Cross-sell through Social Networks*. To appear in European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), 2012.

Challenges in Viral Marketing

- Propagation of influence is a probabilistic process
- The number of individuals in the social network that are getting influenced by the initial trend setters is an *expected quantity*
- Viral marketing for single or multiple products
- There can possibly exists certain types of dependencies among these products

Motivating Example 1: Viral Marketing

- Social networks play a key role for the spread of an innovation or technology
- We would like to market a new product that we hope will be adopted by a large fraction of the network
- Which set of the individuals should we target for?
- Idea is to initially target a few influential individuals in the network who will recommend the product to other friends, and so on
- A natural question is to find a target set of desired cardinality consisting of influential nodes to maximize the volume of the information cascade

Motivating Example 2: Weblogs

- In the domain of weblogs, bloggers publish posts and use hyper-links to refer to other posts and content on the web
- Possible to observe the spread of information in the blogosphere, as each post is time stamped
- In this setting, our goal is to select a small set of blogs (to read) which link to most of the stories that propagate over the blogosphere

Models for Diffusion of Information

- Linear Thresholds Model
- Independent cascade model,
- Decreasing cascade model, etc.

Models for Diffusion of Information

- **Linear threshold model**
- Independent cascade model,
- Decreasing cascade model, etc.

Linear Thresholds Model

- Call a node active if it has adopted the information
- Initially every node is inactive
- Let us consider a node i and represent its neighbors by the set $N(i)$
- Node i is influenced by a neighbor node j according to a weight w_{ij}. These weights are normalized in such a way that

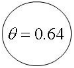
$$\sum_{j \in N(i)} w_{ij} \leq 1.$$

- Further each node i chooses a threshold, say θ_i, uniformly at random from the interval [0,1]
- This threshold represents the weighted fraction of node i's neighbors that must become active in order for node i to become active

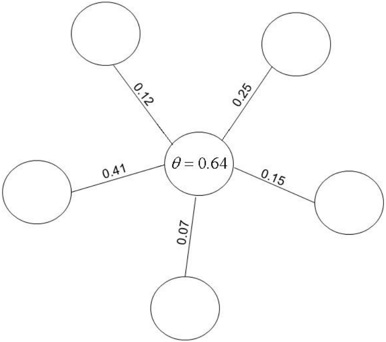
Given a random choice of thresholds and an initial set (call it S) of active nodes, the diffusion process propagates as follows:

- in time step t, all nodes that were active in step $(t-1)$ remain active
- we activate every node i for which the total weight of its active neighbors is at least θ_i
- if $A(i)$ is assumed to be the set of active neighbors of node i, then i gets activated if
$$\sum_{j \in A(i)} w_{ij} \geq \theta_i.$$
- This process stops when there is no new active node in a particular time interval

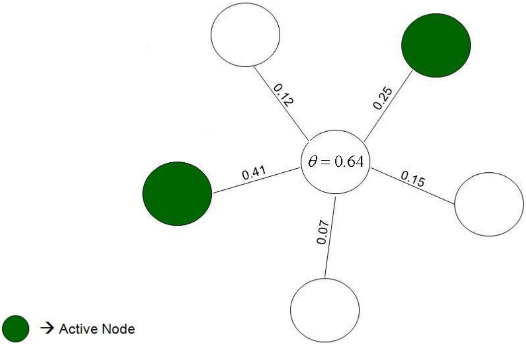
Linear Thresholds Model: An Example



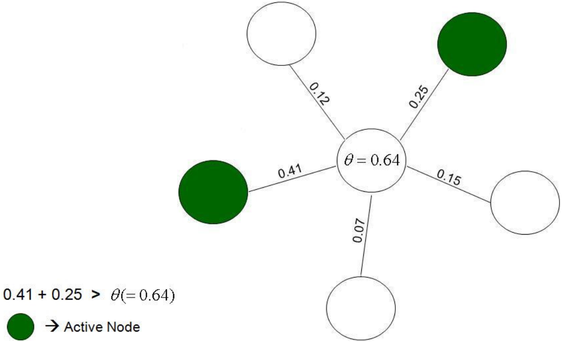
Linear Thresholds Model: An Example



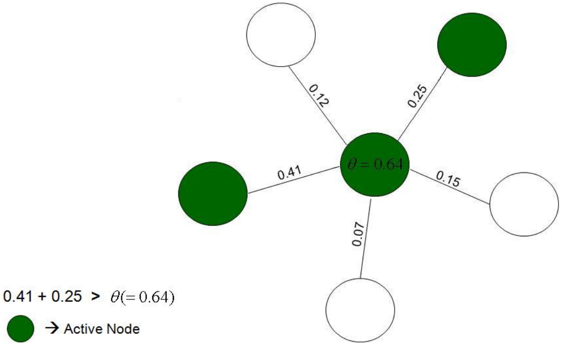
Linear Thresholds Model: An Example



Linear Thresholds Model: An Example



Linear Thresholds Model: An Example



Top-k Nodes Problem

- *Objective function ($\sigma(.)$):* Expected number of active nodes at the end of the diffusion process
- If S is the initial set of target nodes, then $\sigma(S)$ is the expected number of active nodes at the end of the diffusion process
- For economic reasons, we want to limit the size of S
- For a given constant k, the top-k nodes problem seeks to find a subset of nodes S of cardinality k that maximizes the expected value of $\sigma(S)$

Applications

- Databases
- Water Distribution Networks
- Blogspace
- Newsgroups
- Virus propagation networks

- R. Akbarinia, F.E. Pacitti, and F.P. Valduriez. Best Position Algorithms for Top-k Queries. In VLDB, 2007.
- J. Leskovec, A. Krause, and C. Guestrin. Cost-effective outbreak detection in networks. In ACM KDD, 2007.
- N. Agarwal, H. Liu, L. Tang, and P.S. Yu. Identifying influential bloggers in a community. In WSDM, 2008.

A Glimpse of State-of-the-Art

- P. Domingos and M. Richardson. Mining the network value of customers. In ACM SIGKDD, 2001.
- D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In SIGKDD, 2003.
  - Show that the optimization problem of selecting most influential nodes is NP-hard problem.
  - Show that this objective function is a sub-modular function.
  - Propose a greedy algorithm that achieves an approximation guarantee of $(1 - \frac{1}{e})$.

A Glimpse of State-of-the-Art (Cont.)

Greedy Algorithm - KKT (2003)

1. Set $A \leftarrow \phi$.
2. **for** $i = 1$ to k **do**
3. Choose a node $n_i \in N \setminus A$ maximizing $\sigma(A \cup \{n_i\})$
4. Set $A \leftarrow A \cup \{n_i\}$.
5. **end for**

- Running time of Greedy Algorithm: $O(knRm)$.

A Glimpse of State-of-the-Art (Cont.)

- J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In ACM SIGKDD, 2007.
- W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In ACM SIGKDD, 2009.
- N. Chen. on the approximability of influence in social networks. In ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 1029-1037, 2008.
- M. Kimura and K. Saito. Tractable models for information diffusion in social networks. In PKDD, 2006.

Research Gaps

- All the existing approximation algorithms are sensitive to the number of initial trend setters (i.e. k)
- All the approximation algorithms crucially depend on the submodularity of the objective function. It is quite possible that the objective function can be non-submodular

A New Approach

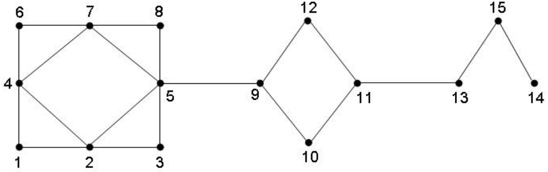
- We present a cooperative game theoretic framework for the top-k nodes problem
- We measure the influential capabilities of the nodes as provided by the Shapley value.
- ShaPley value based discovery of Influential Nodes (SPIN):
  1. Ranking the nodes,
  2. Choosing the top-k nodes from the ranking order.
- Advantages of SPIN:
  1. Quality of solution is same as that of popular benchmark approximation algorithms,
  2. Works well for both sub-modular and non-submodular objective functions,
  3. Running time is independent of the value of k.

Rank-List Construction

1. Let π_j be the j-th permutation in $\hat{\Omega}$ and R be repetitions.
2. Set $MC[i] \leftarrow 0$, for $i = 1, 2, \ldots, n$.
3. **for** $j = 1$ to t **do**
4.   Set $temp[i] \leftarrow 0$, for $i = 1, 2, \ldots, n$.
5.   **for** $r = 1$ to R, **do**
6.   assign random thresholds to nodes;
7.   **for** $i = 1$ to n, **do**
8. $temp[i] \leftarrow temp[i] + v(S_i(\pi_j) \cup \{i\}) - v(S_i(\pi_j))$
9.   **for** $i = 1$ to n, **do**
10. $MC[i] \leftarrow temp[i]/R$;
11. **for** $i = 1$ to n, **do**
12.   compute $\Phi[i] \leftarrow \frac{MC[i]}{t}$
13. Sort nodes based on the average marginal contributions of the nodes

Choosing Top-k Nodes

1. Naive approach is to choose the first k in the RankList[] as the top-k nodes.
2. *Drawback:* Nodes may be clustered.
3. RankList[]=$\{$5,4,2,7,11,15,9,13,12,10,6,14,3,1,8$\}$.
4. Top 4 nodes, namely $\{5, 4, 2, 7\}$, are clustered.
5. Choose nodes:
   - rank order of the nodes
   - spread over the network



k value	*Greedy Algorithm*	*Shapley Value Algorithm*	*MDH based Algorithm*	*HCH*
1	4	4	4	2
2	8	7	7	4
3	10	10	8	6
4	12	12	8	7
5	13	13	10	8
6	14	14	13	8
7	15	15	13	8
8	15	15	13	8
9	15	15	13	10
10	15	15	13	11
11	15	15	13	13
12	15	15	13	13
13	15	15	14	14
14	15	15	15	15
15	15	15	15	15

Running Time of SPIN

- Overall running time of SPIN is $O(t(n+m)R + n\,log(n) + kn + kRm)$ where t is a polynomial in n.
- For all practical graphs (or real world graphs), it is reasonable to assume that $n < m$. With this, the overall running time of the SPIN is $O(tmR)$ where t is a polynomial in n.

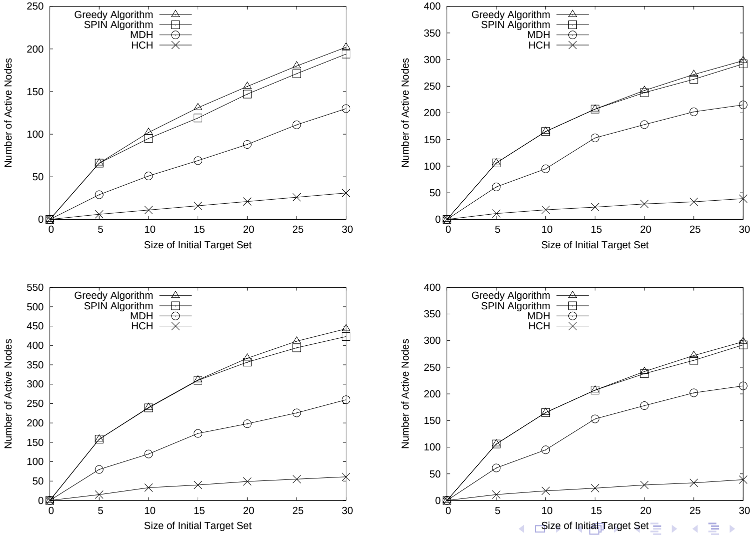
Experimental Results: Data Sets

- Random Graphs
  - Sparse Random Graphs
  - Scale-free Networks (Preferential Attachment Model)
- Real World Graphs
  - Co-authorship networks,
  - Networks about co-purchasing patterns,
  - Friendship networks, etc.

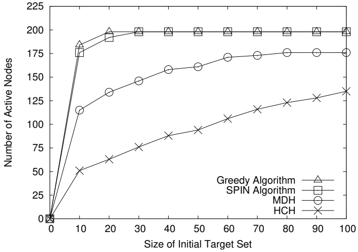
Experimental Results: Data Sets

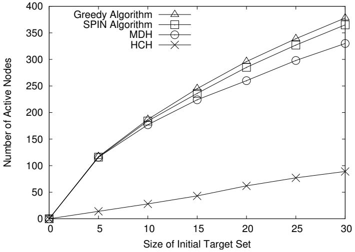
Dataset	Number of Nodes	Number of Edges
Sparse Random Graph	500	5000 (approx.)
Scale-free Graph	500	1250 (approx.)
Political Books	105	441
Jazz	198	2742
Celegans	306	2345
NIPS	1061	4160
Netscience	1589	2742
HEP	10748	52992

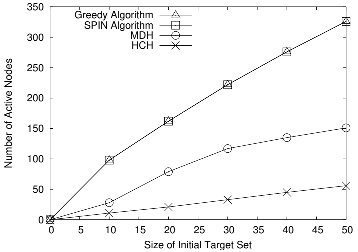
Experiments: Random Graphs

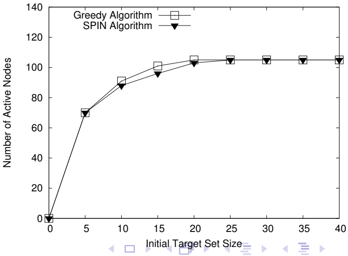


Experiments: Real World Graphs

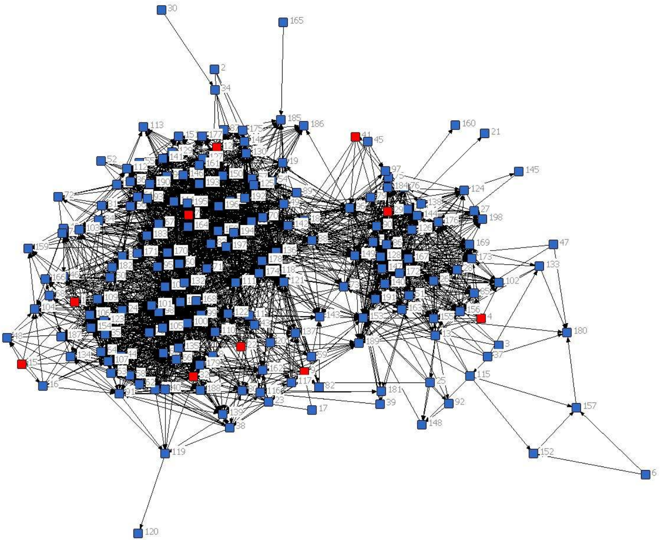




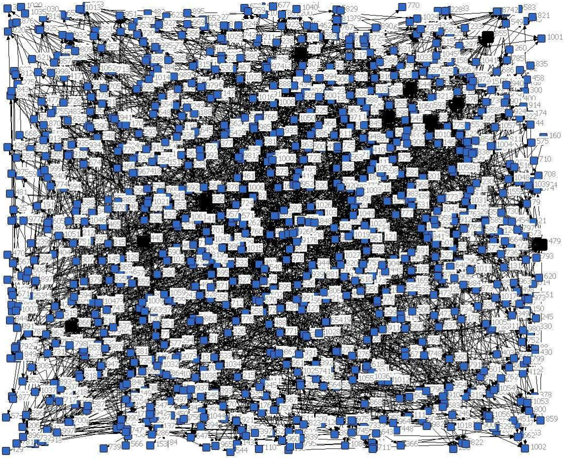




Top-10 Nodes in Jazz Dataset



Top-10 Nodes in NIPS Co-Authorship Data Set



this symbol represents influential node

Running Times: SPIN vs KKT

Dataset	Nodes	SPIN (MIN)	KKT (MIN)	Speed-up
Random graph ($p = 0.005$)	500	13.9	824.93	59
Random graph ($p = 0.01$)	500	14.8	1123.16	75
Random graph ($p = 0.02$)	500	16.3	1302.46	79
Political Books	105	0.89	44.64	50
Jazz	198	1.1	366	332
Celegans	306	14.02	901	64
NIPS	1062	15.2	7201.54	473
Network-Science	1589	28.25	8539.48	302

Table: Speedup of the SPIN algorithm to find Top-30 nodes on various datasets compared to that of KKT algorithm

Running Times: SPIN vs KKT

Top-k Nodes	Running Time (in MINUTES) SPIN Algorithm	KKT Algorithm	LKG Algorithm	Speed-up of SPIN over KKT
$k = 10$	28.04	1341.29	77.07	47
$k = 20$	28.09	4297.02	79.75	152
$k = 30$	28.13	8539.48	85.04	302
$k = 40$	28.18	13949.9	90.33	493
$k = 50$	28.25	20411.1	99.03	722

Table: Running times of the SPIN, KKT, and LKG algorithms on the Netscience data set ($n = 1589$) to determine top-k nodes where $k = 10, 20, 30, 40, 50$ and the speed up of the SPIN algorithm over the KKT algorithm

Next Part of the Presentation

Viral Marketing: Extensions

- Viral Marketing with both Positive and Negative Opinions
- Viral Marketing with Competing Companies
- Viral Marketing with Multiple Independent Products
- Viral Marketing with Cross-sell of Products
- Many more . . .

Viral Marketing with Negative Opinions

- Often not only positive opinions about the products, but also negative opinions may emerge and propagate over the social network.
- How to choose the initial seeds for viral marketing in the presence of both positive and negative opinions?
- W. Chen, A. Collins, R. Cummings, T. Ke, Z. Liu, D. Rincon, X. Sun, Y. Wang, W. Wei, and Y. Yuan. Influence maximization in social networks when negative opinions may emerge and propagate. In SDM 2011.

Viral Marketing: Competing Companies

- A company may introduce some product into the market when certain other company already has introduced a competing product into the market
- How to choose the initial seeds for viral marketing of products in the presence of competing products already present in the market?
- X. He, G. Song, W. Chen, and Q. Jiang. Influence blocking maximization in social networks under the competitive linear threshold model. In SDM, 2012.
- T. Carnes, C. Nagarajan, S. Wild, A. van Zuylen. Maximizing influence in a competitive social network: a followers perspective. In the Proceedings of the 9th International Conference on Electronic Commerce (ICEC), pages 351360, 2007.

Viral Marketing with Multiple Independent Products

- Often companies introduce *multiple independent products* into the market
- Need to satisfy certain domain specific constraints while choosing the initial seeds for each of these products (Eg. no individual should be part of the set of initial seeds for more than certain number of products
- The procedure for selecting the initial seeds for each of the independent products is very different from that of viral marketing with single product
- S. Datta, A. Majumder, N. Shrivastava. Viral marketing for multiple products. In: Proceedings of IEEE ICDM, pages 118127, 2010

Viral Marketing with Cross-sell and Up-Sell of Products

- Often products exhibit relationships among theseselves
- Products may exhibit cross-sell relationship
- Products may exhibit up-sell relationship

Next Part of the Presentation

1. Viral Marketing with Single Product
2. Viral Marketing - Extensions
3. **Viral Marketing with Product Cross-sell**
4. Summary and Conclusions

Product Cross-sell

- *Product Cross-sell Phenomenon*
  - First Variant of Cross-Sell: The use or need for any product in P2 arises only after buying some product in P1.
  - Second Variant of Cross-Sell: An individual can buy a product in P2 before buying any product in P1. However, the possibility of buying products in P2 increases after buying certain products in P1.
- *Product Specific Costs and Benefits*
  - There is a cost associated with each product in order to provide promotional offers (or discounts) to each of the initial seeds
  - There is a benefit associated with each product when it is sold
- *Budget Constraint*
  - A company often has budget constrains and hence the initial seeds have to be chosen within a given budget.

Our Results

- Let B be the budget, c_m be the minimum cost of providing a free sample of any product and c_M be the maximum cost of providing a free sample of any product
- Proposed new information propagation models for product cross-sell
  - Linear Thread with Cross-sell of Products,
  - Modified Linear Thread with Cross-sell of Products.

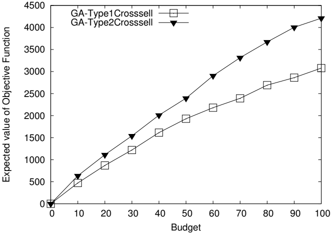
Problem Type	**Proposed Model**	**Approximation Ratio**
Must-Buy Problem	LT-CP Model	$\frac{Bc_m}{B(c_M+c_m)+c_M c_m}$
May-Buy Problem	Modified LT-CP Model	$\frac{Bc_m}{B(c_M+c_m)+c_M c_m}$

Experimental Results: Data Sets

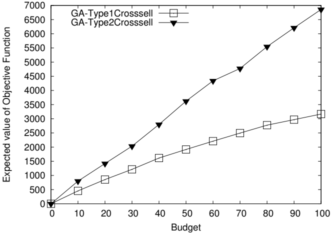
Data Set	Number of Nodes	Number of Edges
WikiVote	7115	103689
HEP	10748	52992
Epinions	75879	508837
Telco Call Data	354700	368175

Comparison of Cross-sell Types

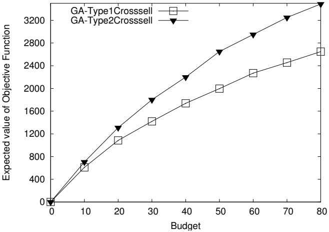
(i)



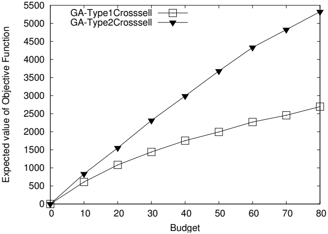
(ii)



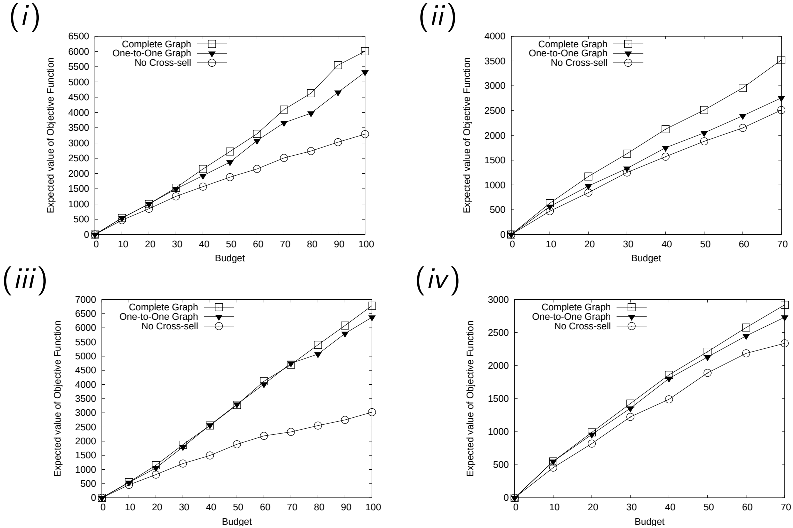
(iii)



(iv)



Cross-sell vs. No-Cross-sell



Next Part of the Presentation

1. Viral Marketing with Single Product
2. Viral Marketing - Extensions
3. Viral Marketing with Product Cross-sell
4. **Summary and Conclusions**

Summary of the Presentation

- We first looked at viral marketing with single product scenario
- Next we considered various extensions to viral marketing
- Finally we looked at viral marketing with product cross-sell

Some Important Latest Developments

- Mingkai Lin, Lintan Sun, Rui Yang, Xusheng Liu, Yajuan Wang, Ding Li, Wenzhong Li, Sanglu Lu. Fair Influence Maximization in Large-scale Social Networks Based on Attribute-aware Reverse Influence Sampling. In JAIR, 76:925-957, 2023.
    - Conventional influence maximization algorithms cause unfair" influence spread among different groups in the population.
- Yandi Li, Haobo Gao, Yunxuan Gao, Jianxiong Guo, Weili Wu. A Survey on Influence Maximization: From an ML-Based Combinatorial Optimization. In ACM Transactions on Knowledge Discovery, 17(9), Article No. 133, pages 150.

Some Important Latest Developments (Cont.)

- Daniel Deutch, Nave Frost, Benny Kimelfeld, Mikal Monet. Computing the Shapley Value of Facts in Query Answering. SIGMOD Conference 2022: 1570-1583.
- Xiaobin Hong, Tong Zhang, Zhen Cui, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jian Yang. Graph Game Embedding. In AAAI, pages 7711-7720, 2021.
- Oliver Biggar and Iman Shames. The graph structure of twoplayer games. In Science Reports, 13:1833, 2023.

Thank You