Influence Diffusion and Optimization in Online Social Networks

Speaker: Dr. De-Nian Yang

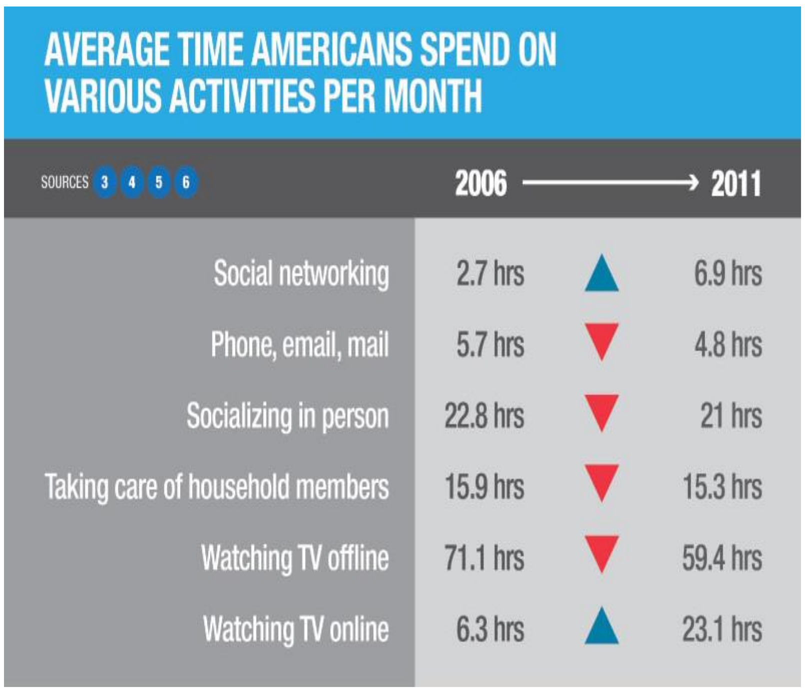
Academia Sinica

Outline

- Importance of Online Social Networks
- Real-World Apps Exploiting Social Influence
- Influence Diffusion Model
- Quick Look on Related Work
- Target-Oriented Social Influence and Active Friending
- When Social Influence Meets Frequent Patterns
- Other related topics

Social Networks Becomes Increasingly Important

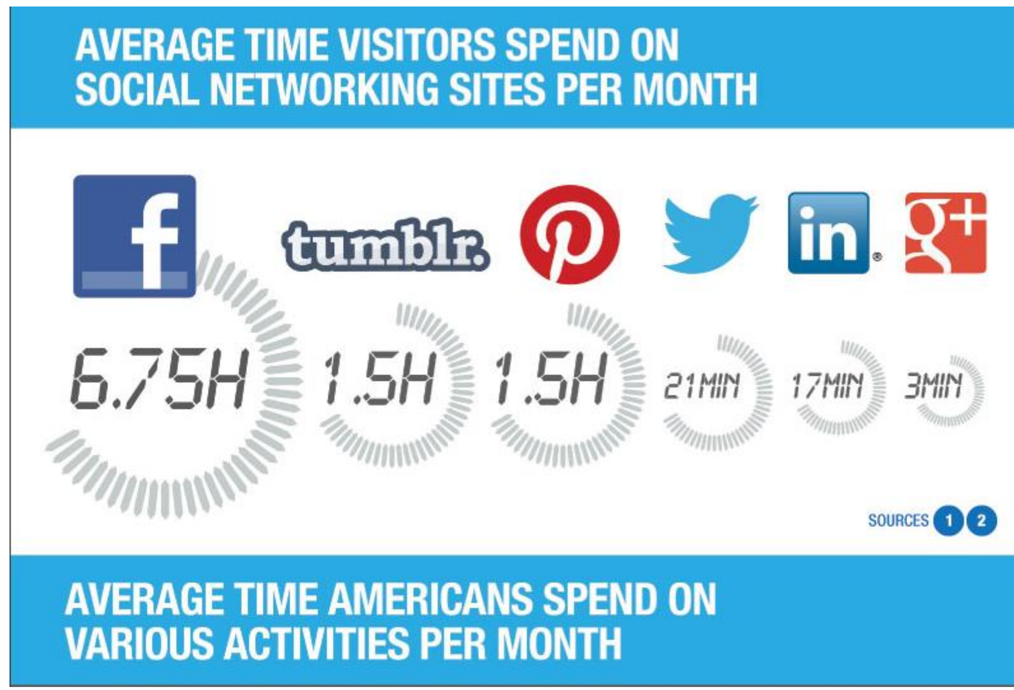
- 1.2 billion users around the world visited social networking websites, accounting for 82 percent of the world's population (**comScore**)
- Nowadays, nearly 1 in every 5 minutes spent online around the world is now spent on social networking websites (**comScore**)



AVERAGE TIME AMERICANS SPEND ON VARIOUS ACTIVITIES PER MONTH

SOURCES 3 4 5 6

	2006		2011
Social networking	2.7 hrs	▲	6.9 hrs
Phone, email, mail	5.7 hrs	▼	4.8 hrs
Socializing in person	22.8 hrs	▼	21 hrs
Taking care of household members	15.9 hrs	▼	15.3 hrs
Watching TV offline	71.1 hrs	▼	59.4 hrs
Watching TV online	6.3 hrs	▲	23.1 hrs



Facebook

- 890 million daily active users on average in December 2014 (**Facebook**)
- 1.19 billion monthly active users who used Facebook mobile products as of December 31, 2014 (**Facebook**)

- More than 201.6 billion friend connections on Facebook at the end of January 2014 (**Facebook**)

- An average of more than 1 billion video views on Facebook per day in June 2014 (**Facebook**)
- About 12 billion messages are sent per day through Facebook in 2014 (**Facebook**)

- Facebook enables advertisers to reach more than one billion people with ads that are relevant and engaging social context (**Facebook**)

Impacts on Business

- Big Business is embracing social media in a big way. The sales of software to run corporate social networks will grow 61% a year and be a $6.4 billion business by 2016 (**USA Today**)
- Yahoo has published a patent detailing how ad charges could be based on a viewer's "social influence". (**BBC News**)
- U.S. Securities and Exchange Commission allows companies to use social media for corporate disclosures (**SEC**)
- Bloomberg integrates live Twitter feeds with financial platform (**Bloomberg**)
  - “When important news is shared on Twitter, traders and investors need to be able to access it, and validate its importance in order to incorporate that information into their decision making process,” said Jean-Paul Zammitt, head of sales and product development for the Bloomberg Professional service.
- Social network is really a great branding tool. (**Bloomberg Businessweek**)
- Brands are building a great presence on social networks and are looking at ways of making it more accessible (**New York Times**)

Impacts on Business

- Twitter speaks, markets listen and fears rise (**New York Times, BBC**)
  - After a Twitter hoax that claimed President Obama was injured in an explosion at the White House. That report caused the Dow Jones industrial average to drop temporarily by 150 points, erasing $136 billion in market value

- Facebook friends could change your cred
  - A handful of tech startups are using social data to determine the risk of lending to people who have a difficult time accessing credit.
- In August 2012, an Italian journalist set up a fake Twitter account for a member of Russia's government and tweeted that the president of Syria had been killed, causing brief fluctuations in the oil markets (**CNN**)

Impacts on Politics (cont.)

- Egyptian Revolution Began on Facebook (**New York Times**)
  - “We Are All Khaled Said” (a page created on Facebook) helped ignite an uprising that led to the resignation of President Hosni Mubarak and the dissolution of the ruling National Democratic Party.

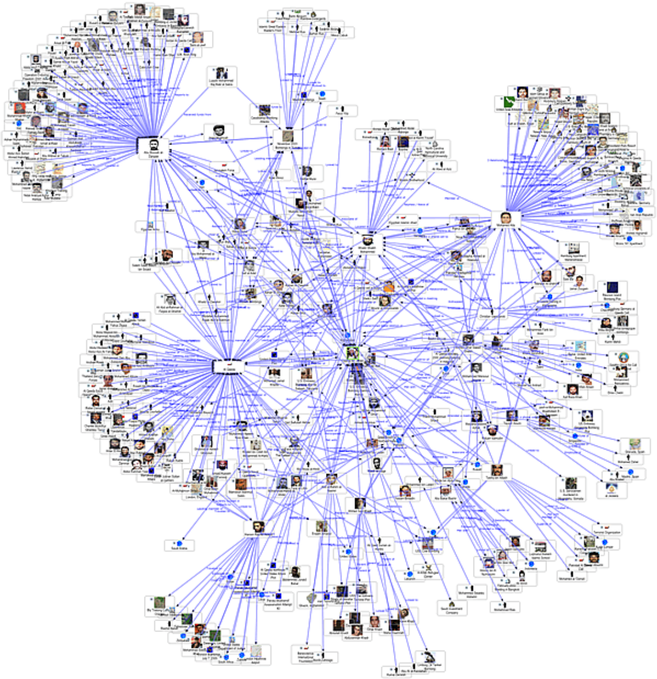
- Tunisian protests fueled by social media networks (**CNN**)

- A tweet doesn't just trigger financial panic, it can also strain diplomatic relations, as the U.S. Embassy in Cairo found out in April when the official Twitter account posted a link to a Daily Show segment critical of Egyptian President Mohammed Morsi (**CNN**)

- In March, someone posing as the U.S. ambassador to Moscow tweeted a criticism of the Russian presidential election process, which was picked up by the news media in Russia before it was revealed as a hoax. The U.S. government responded with official statements in both incidents (**CNN**)

Social Graph

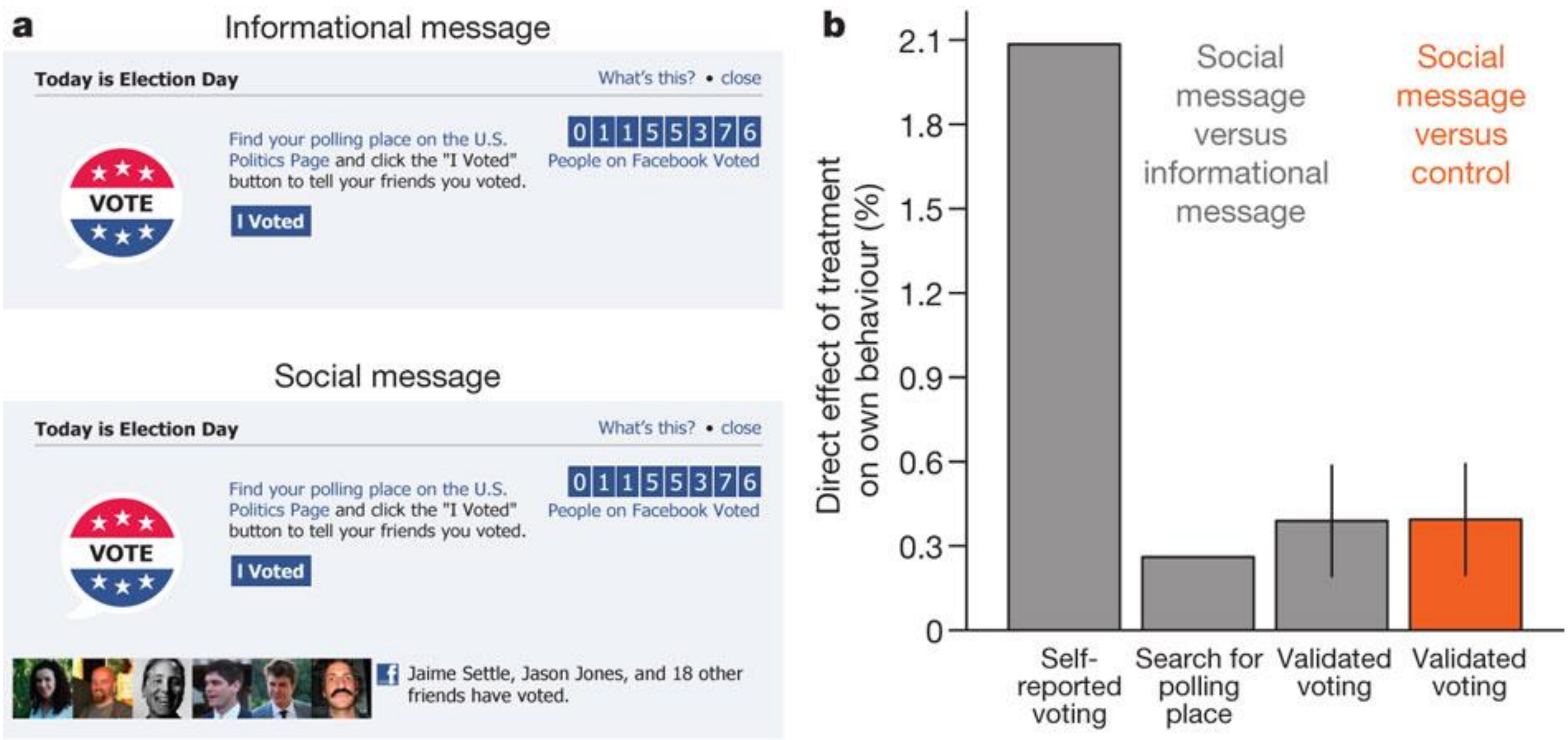
- Previous works analyze the structural properties of social networks (**Nature 1998, Science 1999, Nature 2000**)



- However, those works do not consider
  - The emergence of online platforms (social media & social networking websites)
  - The social influence between users thorough the platforms

Social Influence

- These results show that the people selection tend to be affected by social influence passing through emerging online social networks (**Nature 2012**)
- However, the above analysis is performed from only the perspective of each person, instead of describing the whole network behavior



Search for people, jobs, companies, and more... Advanced 10

Home | Profile | Connections | Jobs | Interests | Business Services | Try Premium for free

Pulse

Worth It?
John A. Byrne

Use the 1-Question Interview to Assess Everything
Lou Adler

How To Diversify Your Life
James Altucher

Lessons from the Ice Bucket Challenge
David Sable

Is the Golf Course a Dying Venue for Dealmakers?
Kevin OBrien

How to Handle Rumors and Gossip in Business or Publi…
Dr. Vivencio (Ven) Ballano

3 Career Lessons I Learned from Making Bad Sandwiches
Kathleen Murphy

Dharmesh Shah Influencer
Dharmesh Shah
Founder and CTO at HubSpot

Follow

The Surprising Brilliance Of The LinkedIn Influencers Program

Aug 6 2013 | 41,586 | 157 | 69

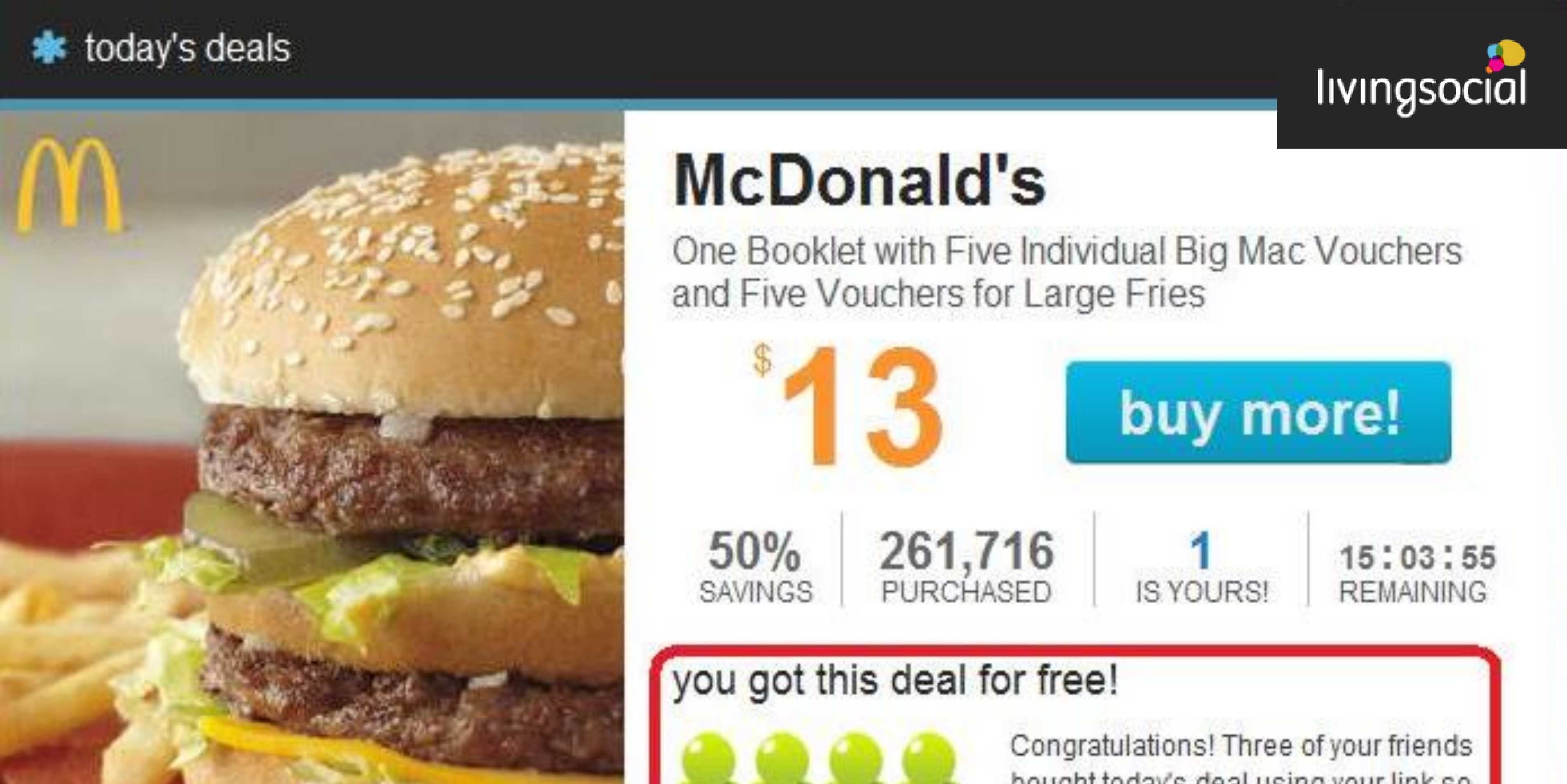
Imagine you have something important to say. (That shouldn't be too hard – we all have something important to say.) You write an article. You post it.

Within 48 hours your post has received over 1 million page views – and from that one article, over 500k people decide to follow you so they can read subsequent posts.

Social Influence in E-Commerce

- Consumers are 71% more likely to make a purchase based on social media referrals (Hubspot)
- 70% of active online adult social networkers shop online (12% more likely than the average adult internet use), and 47% of them are more likely to be heavy spenders (Nielsen).

- Digital marketing agency ODM Group
  - 74% of consumers rely on social networks to guide purchase decisions
  - On Twitter, 53% of consumers recommend companies or products in their tweets. Of those, 48% follow through with the intent to buy that product or service.
  - The most effective platforms in terms of mobilizing consumers to talk about products are Facebook with 86%, followed by Twitter at 65%, blogs and reviews are tied at 55%, and videos come in last with 50%



How does the Me+3 promotion work?

Because we adore you so very much, whenever you purchase a deal, we'll provide you with a referral link that you can send to your friends, telling them about the great deal you just purchased. If three of your friends purchase the deal by using your referral link… wait for it… your deal is free!

活動辦法

免費試駕體驗 | 提交試駕文

免費試駕體驗活動辦法

報名日期 2014/4/3~2014/7/31

試駕日期 2014/4/3~2014/7/31(7/1-7/31限平日試駕，詳情請見試駕體驗券使用規則)

試駕流程

STEP 1 至經銷商端出示證件、填寫問券、繳交保證金

STEP 2 報名完成。同步將資料傳送給租車公司核對

STEP 3 租車公司聯繫消費者，完成預約試乘相關手續

STEP 4 消費者試駕ECOSPORT24小時

STEP 5 試駕完成，前往租車中心還車

STEP 6 分享試駕心得至各大網站

STEP 7 進入活動網頁回報，並填寫後測問卷

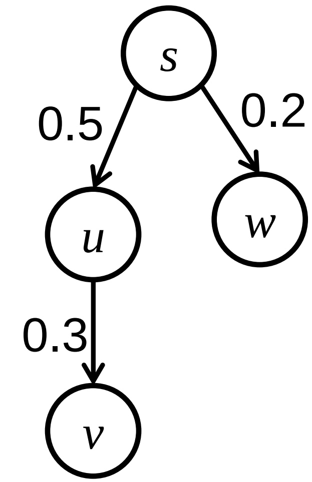
STEP 8 上傳心得網址，審核確認後即可退回保證金

Viral Marketing – Problem Formulation (KDD'03)

- The world-of-mouth or viral marketing is an effective marketing strategy
  - Based on trust among individuals' close social cycles
  - Carefully select seeds to maximize the number of influenced users (spread)
- Influence maximization problem
  - Given
    - A social network $G(V,E)$
    - A constant k
  - Selects k users in G to maximize the spread
    - Social influence model: determine whether a user is influenced or not

Independent Cascade Model (KDD'03)

- Given the social network $G(V,E)$
  - $w(x,y)$: influence weight from x to y
- At step 0, only seed node s is activated (influenced)
- For every node x activated at step t, it is given a chance to activate its neighbors y with probability $w(x,y)$. If succeeds, y is activated at step $(t+1)$
- The process ends if no new activated nodes are generated

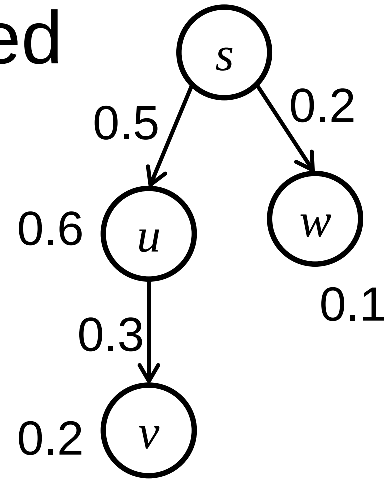


Linear Threshold Model (KDD'03)

- Given the social network $G(V,E)$
  - $w(x,y)$ influence weight from x to y
  - θ_v is randomly uniformly assigned from [0,1]
- At step 0, only seed nodes s are activated
- A node y is activated if

$$\sum_{x \text{ is an active neighbor of } y} w(x,y) \geq \theta_y$$

- The process ends if no new activated nodes are generated



Other Factors – Personal Preference (WSDM'12)

- Personal preference
  - A user' preference on the product also determines whether she will recommend the product to her friends
- Problem: select k seeds to maximize the spread
- LT-C Model
  - A user v is activated if

$$f_v(A) = \frac{\sum_{u \in A} w_{u,v}(r_{u,i} - r_{\min})}{r_{\max} - r_{\min}}$$

    - $r_{u,i}$: the rating (preference) of user u on product i
    - r_{min}/r_{max}: the minimal/maximal rating of all users
    - A: the set active neighbors of v

Other Factors – Negative Opinions (SDM'11)

- Negative opinions
  - Negative opinions about a product may affect users' decision to adopt this product
- Problem: select k seeds to maximize the positive spread while some users may be influenced by negative opinions
- IC-N model
  - Node status: *inactive*, +*active*, -*active*
  - If a user v is activated by a -*active* neighbor, v is -*active*
  - If v is activated by a +*active* neighbor
    - v becomes +*active* with probability q
    - v becomes -*active* with probability 1-q

Other Factors – Existing Competitors (SDM'12)

- Competitive social influence
  - The existing of competitors affects users' decision
- Influence Blocking Problem: given the seeds of competitors S_C and a constant k, select k seeds to minimize the spread of competitors
  - Only one competitor is considered
- Competitive Linear Threshold Model
  - Node status: *inactive*, *active*, *active-c*
  - A node v becomes *active* if

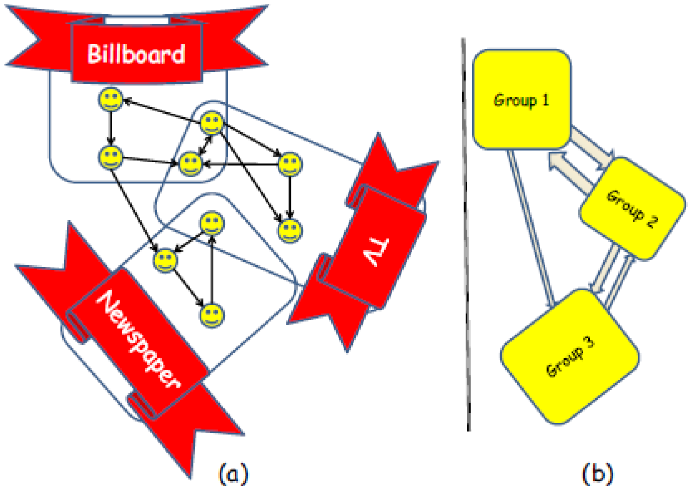
$$\sum_{x \text{ is } active \text{ and a neighbor of } v} w(x,v) \geq \theta_v$$

  - A node v becomes *active-c* if

$$\sum_{x \text{ is } active-c \text{ and a neighbor of } v} w^c(x,v) \geq \theta_v^c$$

Other Factors – Media (KDD'13)

- Consider the influence from media, e.g., TV, billboards, newspaper
- Problem: to select the l influential media to maximize the spread
  - Given
    - the social network
    - the target audience (group) of each medium
  - Selecting a medium will influence a portion of its target audience
  - Social influence is considered in group-level



Other Factors – Conformity (KDD'13)

- Conformity: a type of social influence from groups that a user belongs to
- Problem:
  - Given
    - Social network G(V,E)
    - Members of groups (i.e., whether a user in a group)
    - User interests/attitudes identified from profiles/posts
  - Study how a user's behavior conforms to her peer friends and the communities (groups) that she belongs to

Learning Influence Probability

- The influence probability may be derived from the actions of users
- An action $(v,a,t) \rightarrow$ a user v performs the action *a* at time t
  - a: rating a movie, adopting a product etc.
- If v performs action a and then her friend u also performs $a \rightarrow v$ influence u to perform a
- v may be influenced by multiple users
  - If more than one friends perform a before u
  - They share the ``credits'' to influence u

Learning Influence Probability (cont.)

- Partial Credit model
  - Assume that the influence probability is static and does not vary with time
  - The credit for given to v for influence u on action a

$$credit_{v,u}(a) = \frac{1}{\sum_{w \in S} I(t_w(a) < t_u(a))}$$

    - S: the set of active neighbors (friends) of u
    - I: indicator function
    - $t_w(a)$ and $t_u(a)$: the time when w and v perform a

Learning Influence Probability (cont.)

– The influence probability of v to influence u

$$p_{v,u} = \frac{\sum_{a \in \mathbf{A}} credit_{v,u}(a)}{A_v}$$

- $\mathbf{A}$: the set of actions
- A_v: the number of actions performed by v

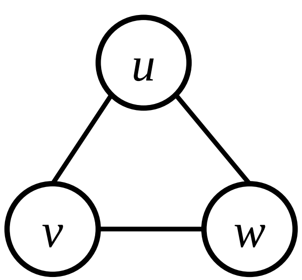
– The joint influence probability to influence u

$$p_u(S) = 1 - \prod_{v \in S} (1 - p_{v,u})$$

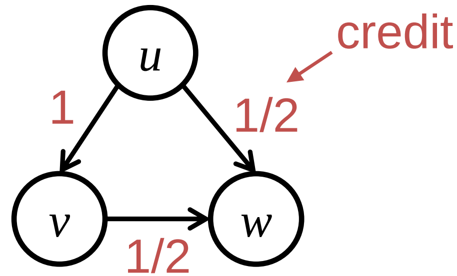
Learning Influence Probability - Example

user	*action*	*time*
u	*a1*	5
v	*a1*	10
w	*a1*	15
v	*a2*	12
w	*a2*	14
w	*a3*	6
u	*a3*	14

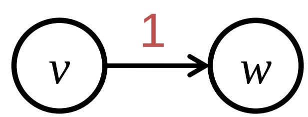
▲Action log



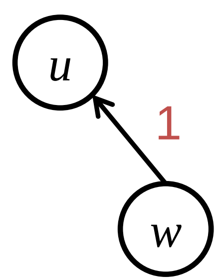
▲ Social graph



▲ Influence propagation of *a1*



▲ Influence propagation of *a2*

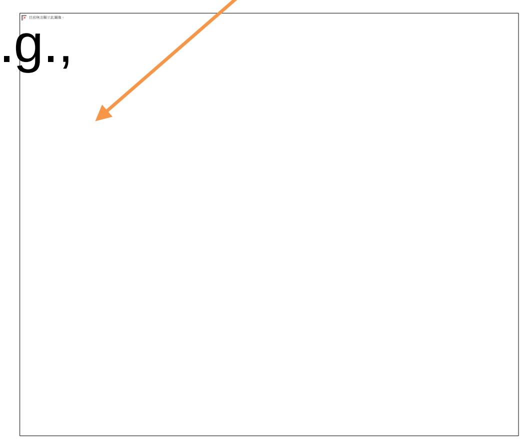


▲ Influence propagation of *a3*

	u	***v***	***w***
u		1/2	1/4
v	0		3/4
w	1/3	0	

▲Derived influence probability

E.g.,



Influence Broadcast

- Influence broadcast problem: given a budget k, select k seeds that maximizes the spread
  - Under IC/LT is NP-Hard
- The spread function is submodular, therefore there is a greedy algorithm with approximation ratio $1\text{-}1/e$
  - Choose the node with the largest marginal effect
- Submodular property, for a function f

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T) \text{ for all } S \subseteq T$$

Finding Broadcast Area is Difficult

- Influence broadcast focuses on the seed selection to maximize the number of influenced nodes
  - Totally k users are selected
  - NP-Hard but approximable within $1\text{-}1/e$

- Given network and seeds, calculating the accurate broadcast area under IC model is #P-Complete
  - Much more challenging
  - Reduction from s-t connectedness

#P-Complete Reduction

- The problem of s-t connectedness in a directed graph is #P-Complete
  - Given a graph $G(V,E)$ and two vertices s and t, count the number of subgraphs in which s is connected to t
  - Counting process equivalent to compute the P (s is connected to t) when each edge has an independent probability of ½ to be connected, ½ to be disconnected
  - P(s is connected to t) = P(t is activated)
- Reduction to the accurate spread calculation
  - Vertex s is the only node in S
  - For each edge, $w(u,v)$=½

#P-Complete Reduction (cont.)

- Compute the spread $I_1 = \sigma_I(S,G)$
- Add a new node *t'* and a directed edge *(t,t')* with *w(t,t')*=1 in *G* to create *G'*
- Compute the spread $I_2 = \sigma_I(S,G) + p(S,t,G) \cdot w(t,t')$
  - $p(S,v,G)$ denotes the probability that *v* is influenced by seedset *S* in *G*
- $I_2 - I_1 = p(S,t,G)$, i.e. the probability *s* is connected to *t*

- Calculating spread accurately is #P-Complete
  - At least as difficult as NP-Complete problems

Linear Threshold v.s. Live-Edge Graph

- Given an influence graph $G(V,E,w)$
  - V: nodes, E: edges, and $w()$: the influence of a edge
- For every node v in V, select at most one of its incoming edge at random
  - The probability that an edge (u,v) is selected is $w(u,v)$
    - The selected edge is *live*, otherwise *blocked*
  - The probability that no incoming edge of v is selected is $1\text{-}\Sigma_u w(u,v)$
  - The live-edge graph R_G is the random graph including V and all live-edges

Linear Threshold v.s. Live-Edge Graph

- Linear Threshold Model
  - Define A_t as the set of active nodes at the end of iteration t. $t=0,1,2\ldots$
  - Consider a node v …
    - It has not become active by the end of iteration t-1,
    - The probability that v becomes active in iteration t is equal to the chance that the weights in $A_t \backslash A_{t-1}$ push v to be activated

$$\frac{\sum_{u \in A_t \backslash A_{t-1}} w(u,v)}{1-\sum_{u \in A_{t-1}} w(u,v)}$$

Linear Threshold v.s. Live-Edge Graph

- Live edge path: A node x ends up active if and only if there is a path consisting of live edges
  - Starting with the active neighbor set A, for each node x with at least one edge from A, determine whether x's live edge comes from A
    - Yes, x is reachable from A
    - No, the live edge of x comes from outside A
  - If the live edge of a node v is unknown, the probability that it is determined in stage t is the chance that it comes from $A_t \backslash A_{t-1}$

$$\frac{\sum_{u \in A_t \backslash A_{t-1}} w(u,v)}{1 - \sum_{u \in A_{t-1}} w(u,v)}$$

Comparison

- Comparison of IC/LT model
  - IC: a node v has a probability $w(u,v)$ to become active if its in-neighbor u becomes active
  - Live-edge graph: at most one of the incoming edge of node v is selected
  - LT: the activation probability of node v is the sum of the influence scores from active neighbors exceeds a threshold

Spread Estimation - Motivation

- Find the influence broadcast area in LT/IC is #P-Hard
  - Run Monte-Carlo simulation many times is required
- Facebook graph has 1.2 billion nodes
- To overcome the inefficiency issue,
  - Estimate the spread based on the DAGs reduced from the original network
  - Perform influence maximization under the estimated graph

Spread Estimation

- Estimate the influence from node *u* to node *v* using the maximum influence path (*MIP*)
  - Other paths from *u* to *v* are eliminated
- Build maximum influence in(out)-arborescence (*MIIA*, *MIOA*) for each node in the network
  - *MIIA(v, θ)*: nodes influence *v* with *MIP* ≥ *θ*

$$MIIA(v,\theta) = \cup_{u\in V, pp(MIP_G(u,v))\geq\theta} MIP_G(u,v)$$

  - *MIOA(v, θ)*: nodes are influenced by *v* with *MIP* ≥ *θ*

$$MIOA(v,\theta) = \cup_{u\in V, pp(MIP_G(v,u))\geq\theta} MIP_G(v,u)$$

  - *MIP*s of those nodes are included in *MIIA*/*MIOA*
  - *θ* is a truncation parameter

Spread Estimation (cont.)

- $ap(u, S, MIIA(v, \theta))$
  - Activation probability of node *u* when the seed set is *S* and the influence is propagated through *MIIA(v, θ)*
  - If *u* is a seed in *S*, *ap(u, S, MIIA(v, θ))= 1*
  - If *u* is not a seed,
    - if *u* has no in-neighbors, *ap(u, S, MIIA(v, θ))=0*
    - Otherwise, $ap(u) = 1 - \prod_{x \in N^{in}(u)} (1 - ap(x) \cdot w(x, u))$

- Final spread
  - $\sigma_M(S) = \sum_{v \in V} ap(v, S, MIIA(v, \theta))$
  - The spread function is submodular and monotone,

Algorithm – Influence Maximization

- Greedy: select the node with the maximum increment of a spread
- Suppose we want to select a new node into *S*
  - Compute the whole spread when adding a candidate $\rightarrow O(t^2)$
  - Improving it to *O(t)* by influence linearity
    - $\Delta ap(v) = \alpha(u,v) \cdot \Delta ap(u) + \beta(u,v)$
    - $\alpha(u,v)$ and $\beta(u,v)$ only depends on $ap(x)$, where x is on the path from u to v; find them at each iteration in *O(t)*
    - $\alpha(u,v)$ is computed recursively
      - $\alpha(u,v) = 1$ for $u = v$
      - $\alpha(u,v) = 0$ if the out-neighbor x of u is a seed
      - Otherwise, $\alpha(u,v) = \alpha(x,v) \cdot w(u,x) \cdot \prod_{u' \in N^{in}(x) \setminus \{u\}} (1 - ap(u') \cdot w(u',x))$

Influence Unicast and Active Friending

- Passive friending (friend recommendation)
  - A user passively selects candidates from the recommendation friend list to send invitations
  - Common in modern social networking websites
- Active friending
  - A user may want take proactive actions to make friends with another person in real life
    - A student fan may like to make friends with the captain in the soccer team
    - A salesperson may want to get acquainted with a high-value potential customer

Active Friending

Active Friending – Motivation (cont.)

- However, no active friending services exist

- A social networking website may recommend suitable candidates *iterative* for a user to assist her in effectively approaching her target
  - With the social topology kept in social networking websites
  - For an active friending initiator s and her target user (the one s wants to meet) t, we aim to find a ``way'' for s to follow to meet t
  - Avoid privacy issues

VIEW

3 COMMENTS

Emerging Technology From the arXiv
March 4, 2013

The Algorithm That Helps You Friend People You Don't Know

Computer scientists have developed an algorithm that uses the structure of a social network to find the best strategy for friending people you don't know

MIT Technology Review

- “The Algorithm That Helps You Friend People You Don’t Know,”

- “Imagine you want to friend an influential person on Facebook who you don’t know and with whom you have no friends in common. How would you go about the task?”

- “One option is simply to send an invitation directly to that person. But without anybody to recommend you, the chances of him or her accepting the invitation are slim.”

Active Friending

- "Researchers Develop Algorithm to Maximize Friendship Acceptance by Strangers on Social Networks,"

ACM TechNews acm.org PHYS.ORG

- "Facebook Friend People You Don't Know, Math Nerds Tell You How,"

THE INQUISITR GOSSIPY BREAKING CELEBRITY NEWS

- "An Algorithm That Helps You Stalk, Er, Meet New People On The Internet,"

POPULAR SCIENCE THE FUTURE NOW

Active Friending – Motivation (cont.)

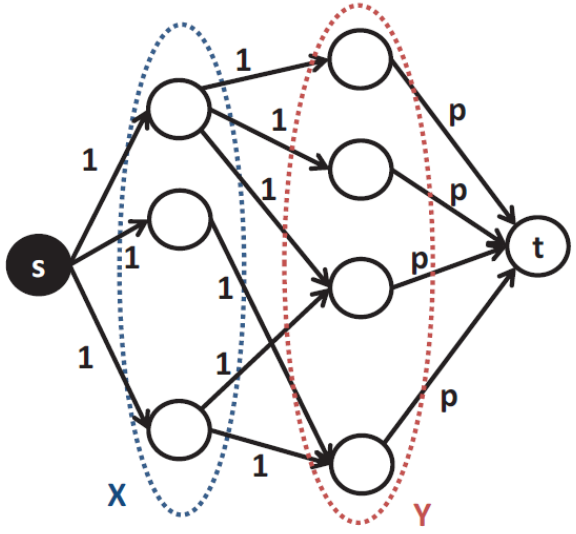
- The number of common friends is important
  - The more common friends, the larger probability to make friend
  - Most existing websites contain the information
  - Having common friends may make a user trust a stranger more

APM Problem for Influence Unicast

- Given
  - A social network $G(V,E)$
  - Source: an active friending initiator s
  - Destination: the target t of s
  - Budget: invitation number constraint r_R
- Acceptance Probability Maximization (APM) problem for influence unicast
  - Aim to select a set R of r_R users
  - Maximize the acceptance probability of t

APM in General Graph

- Calculating acceptance probability of given R is #P-hard
  - By a reduction from s-t connectedness problem
- APM in general graphs is NP-hard
  - By a reduction from set cover problem
  - Set cover problem
    - Aiming to find a subset of X covering all elements in Y



APM in General Graph (cont.)

- Influence broadcast in general graphs is submodular
- APM in general graphs is not submodular
  - A counter example

Approximate Acceptance Probability

- Adopt MIA model to estimate the acceptance probability
  - Approximate the social influence by maximum influence path (MIP)
  - Create a maximum influence in-arborescence $MIIA(t, \theta)$ to estimate the influence to t
    - θ: to truncate users with too small influence to reach t

Approximate Acceptance Probability

- Acceptance probability *ap(v)* of a user *v*
  - If *v* is a friend of *s*, *ap(v)* is 1
  - If *s* does not send an invitation to *v*, *ap(v)* is 0
  - Otherwise, *ap(v)* is derived according to the acceptance probability of its friends

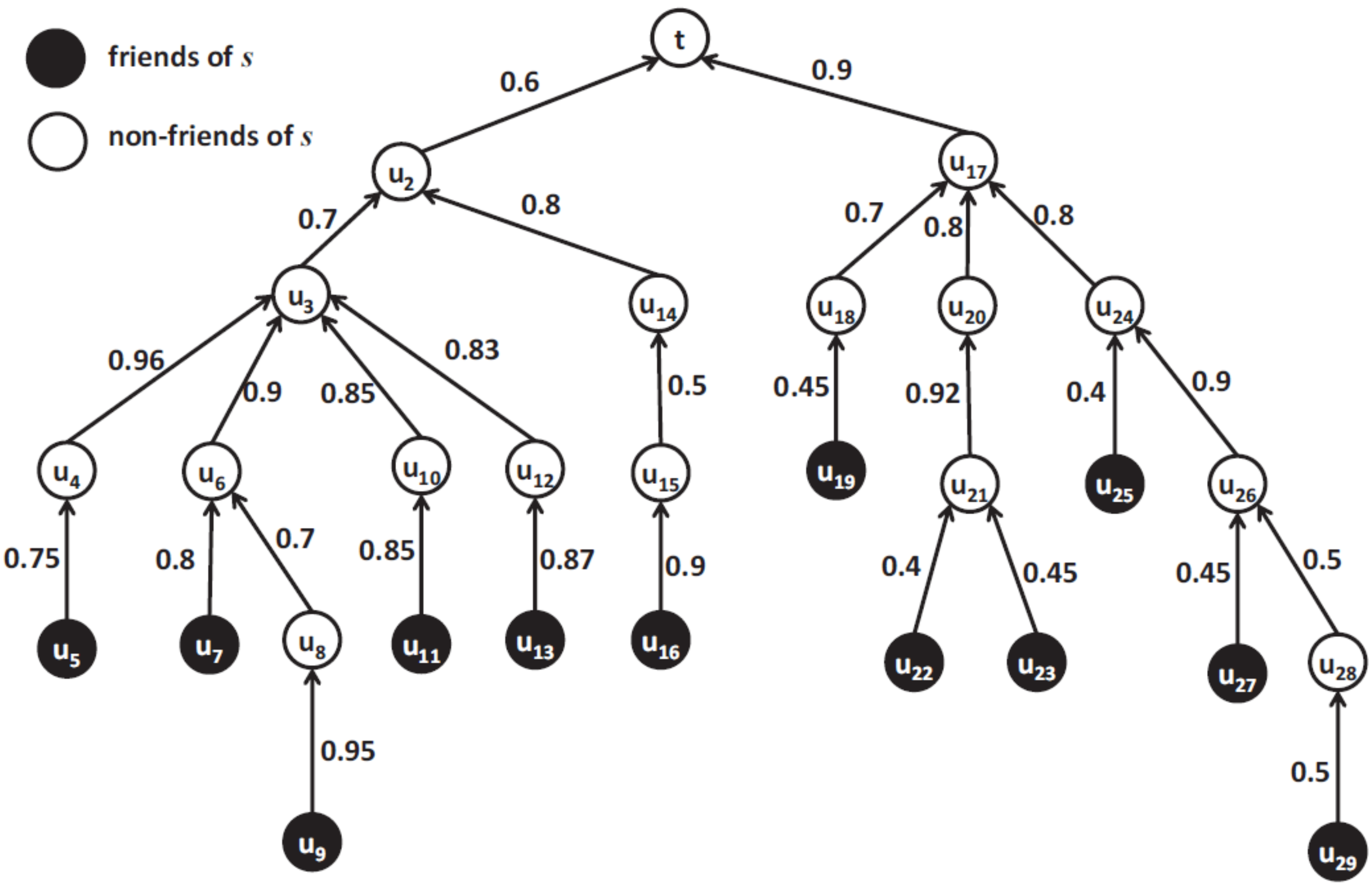
DEFINITION 2. *The acceptance probability for an invitation of a node v in $MIIA(t,\theta)$ is* $ap(v, S, R, MIIA(t,\theta))) =$

$$\begin{cases} 1, \textit{ if } v \in S \\ 0, \textit{ if } v \notin R \textit{ or } N^{in}(v) = \emptyset \\ 1 - \prod_{u \in N^{in}(v), u \in R} (1 - ap(u, S, R, MIIA(t,\theta)) \cdot w_{u,v}) \\ , \textit{ otherwise} \end{cases}$$

where $N^{in}(v)$ *is the set of in-neighbors of* v.

S: friends of *s*   *R*: uses that *s* send a invitation

Example



friends of s
non-friends of s
t
0.6
0.9
u_2
u_{17}
0.7
0.8
0.7
0.8
0.8
u_3
u_{14}
u_{18}
u_{20}
u_{24}
0.96
0.9
0.85
0.83
0.5
0.45
0.92
0.4
0.9
u_4
u_6
u_{10}
u_{12}
u_{15}
u_{19}
u_{21}
u_{25}
u_{26}
0.75
0.8
0.7
0.85
0.87
0.9
0.4
0.45
0.45
0.5
u_5
u_7
u_8
u_{11}
u_{13}
u_{16}
u_{22}
u_{23}
u_{27}
u_{28}
0.95
0.5
u_9
u_{29}

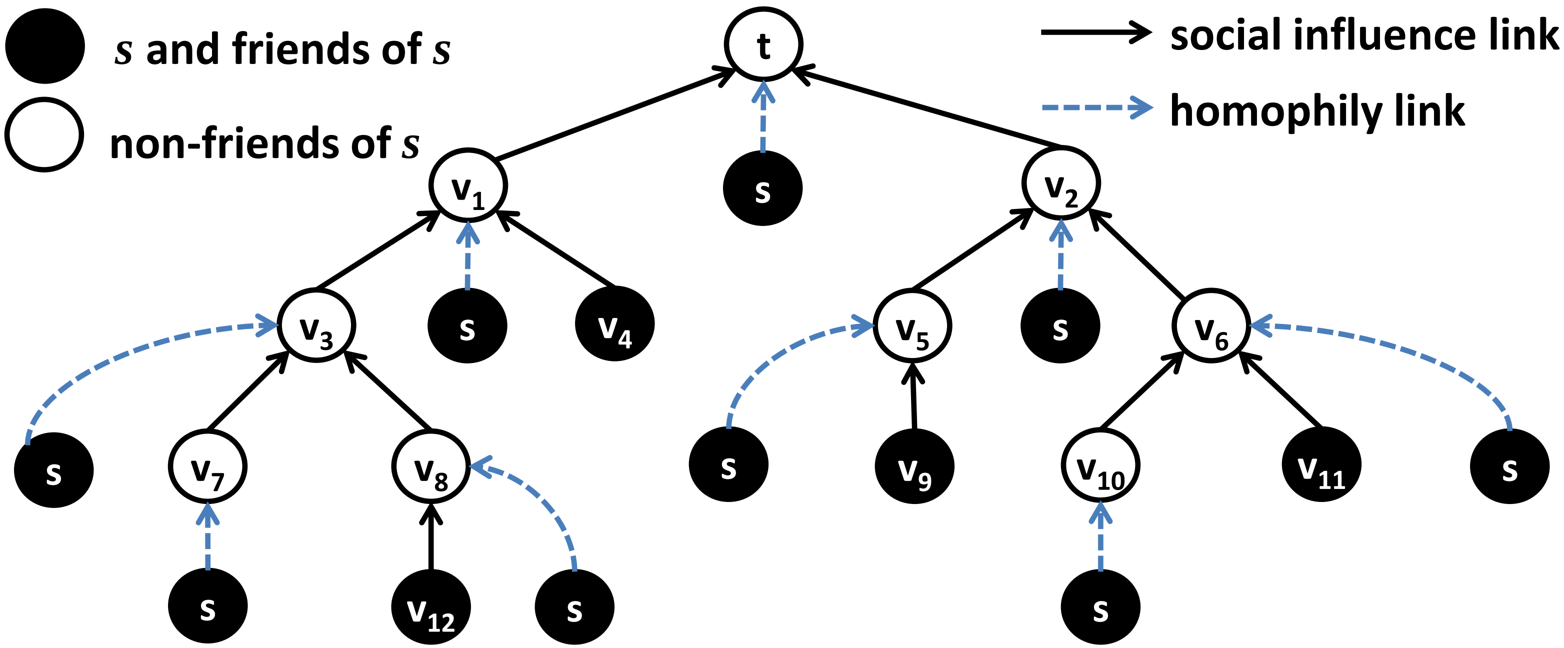
Approximate Acceptance Probability

- Not only the social influence affects the acceptance probability

- Homophily factor captures the effect of characteristics of individuals
  - Similarity and preference etc.

- Extend MIA to consider the homophily factor by duplicating s

Approximate Acceptance Probability (cont.)

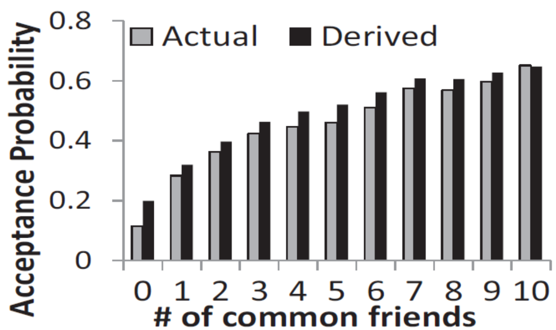


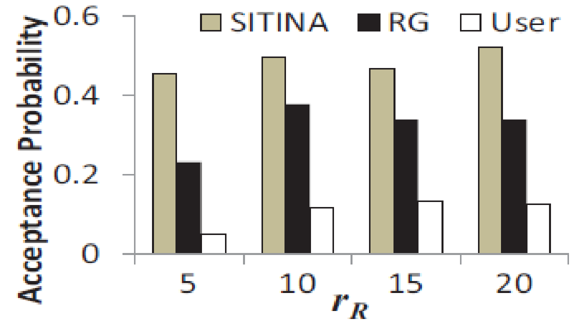
Addressing APM

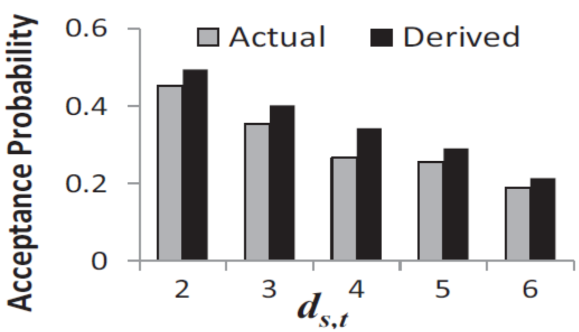
- Propose an algorithm SITINA to solve APM in polynomial time
  - Distribute the (r-1) invitations to its in-neighbors
  - Try all possible invitations takes exponential time
  - Ordering the in-neighbors and storing the maximum acceptance probability when x invitations are sent to the first k in-neighbors
    - To avoid too many combinations

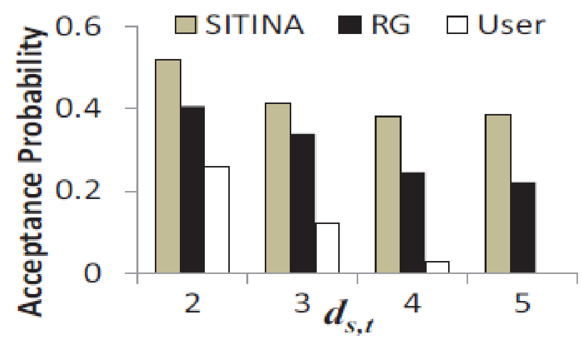
Implementation and User Study

- User study of 169 people (manually coordination)
  - Acceptance probability estimation is accurate
  - APM solutions are much better



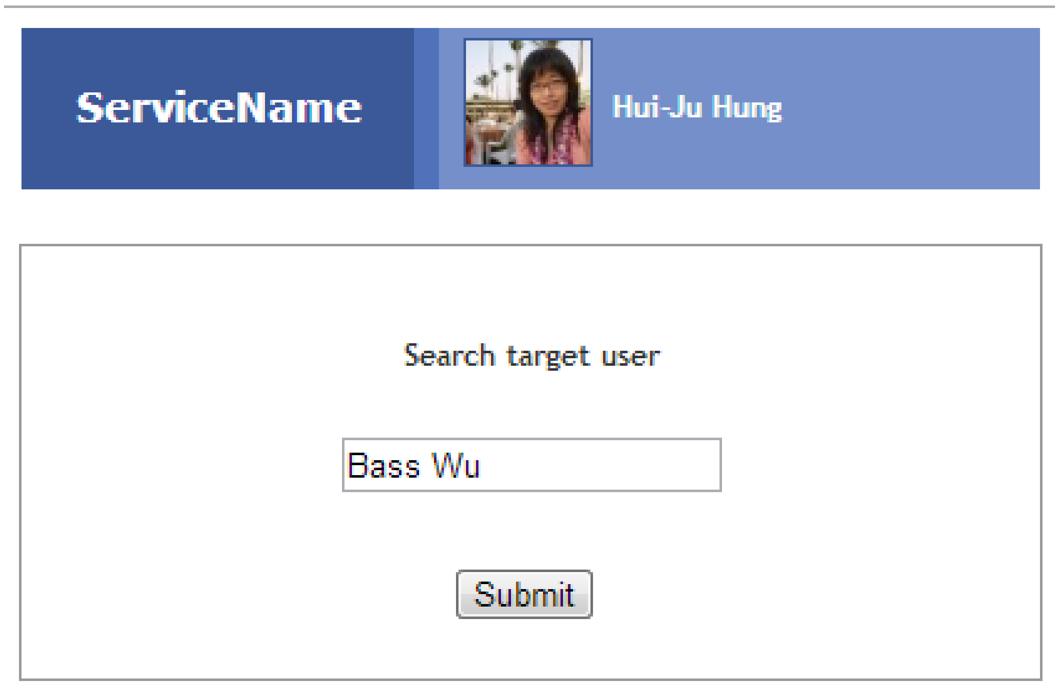






Facebook Implementation

- Search a targeted user

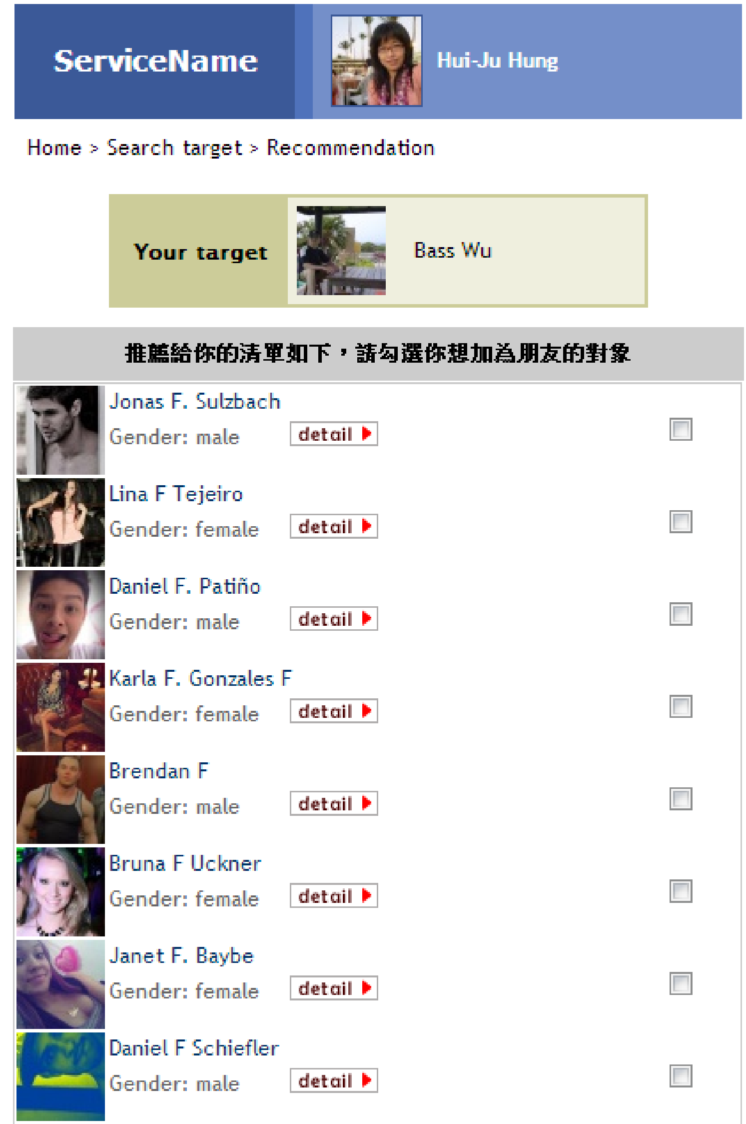


Facebook Implementation

- Select the targeted user from the search results

Facebook Implementation

- Recommended next hop to the targeted user



A Closer Look on Homophily Inference Propagation

- Associations among items in transactions are widely adopted in online e-commerce stores

Figures: http://www.amazon.com/One-Day-Vintage-Contemporaries-Original/dp/0307474712/

More Sophisticated Model on Decision Making

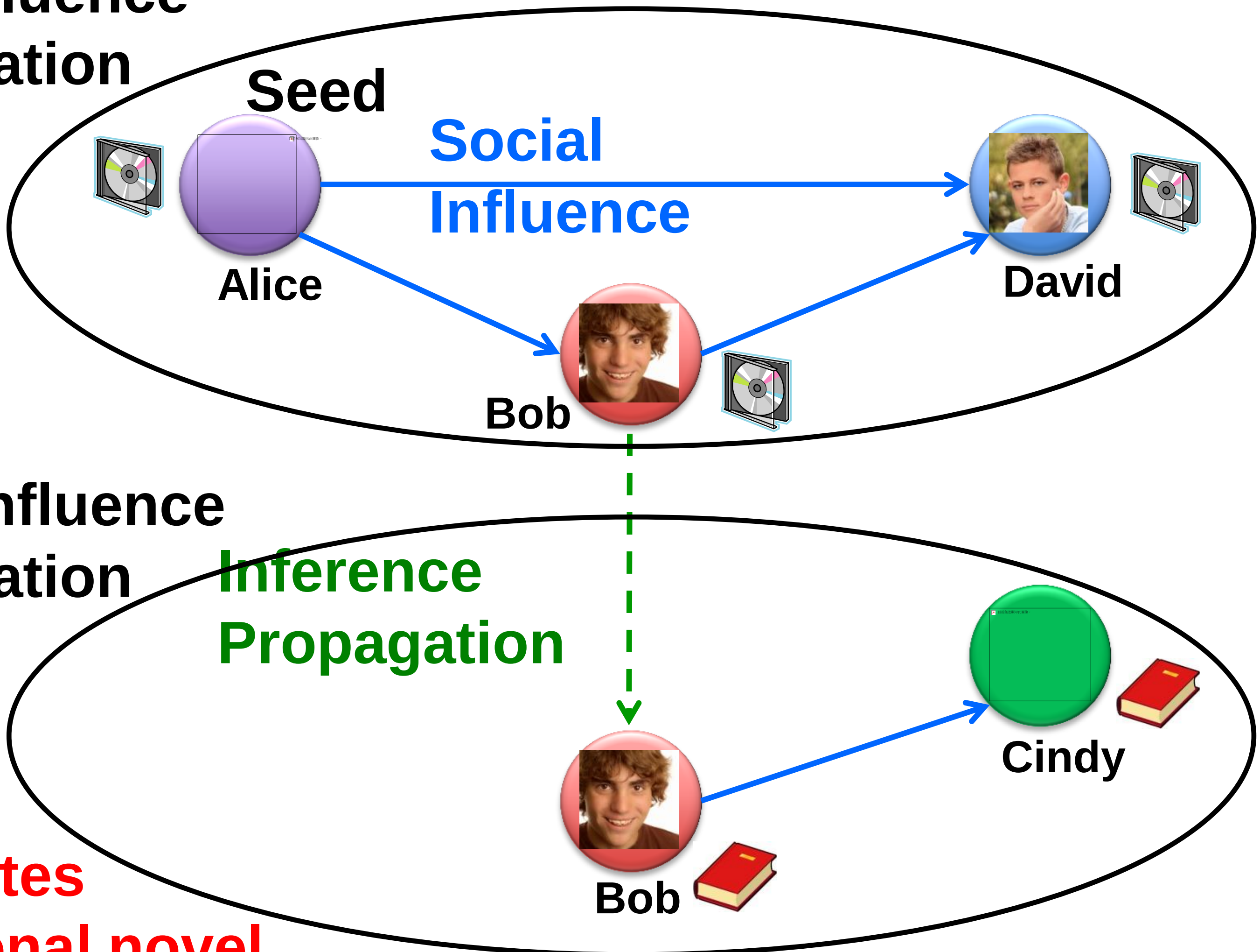
- Previous works
  - Horizontal: social influence propagation (for each product) e.g. Alice -> Bob -> Cindy
  - Vertical: homophily inference propagation (for each person), e.g. DVD-> book

- In reality : horizontal + vertical, i.e., chain effect
  - Alice (DVD) -> Bob(DVD) -> Bob(novel) -> Cindy (novel)
  - Bob buys the novel thanks to the DVD, but he would not buy DVD if not influenced by Alice
  - Thanks to item inference (DVD->novel), **implicit** novel influence broadcast is created by **explicit** DVD influence spread

When Social Influence Meets Frequent Patterns

DVD Influence Propagation

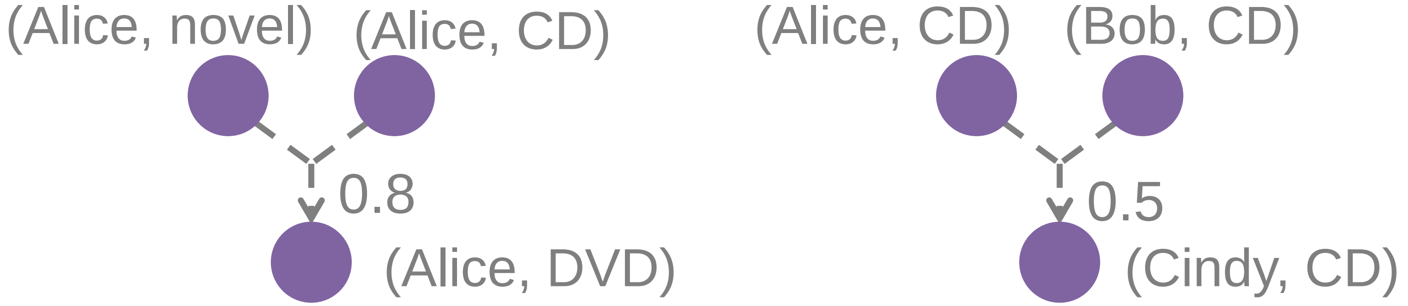
Novel Influence Propagation



DVD seed creates an additional novel broadcast area!

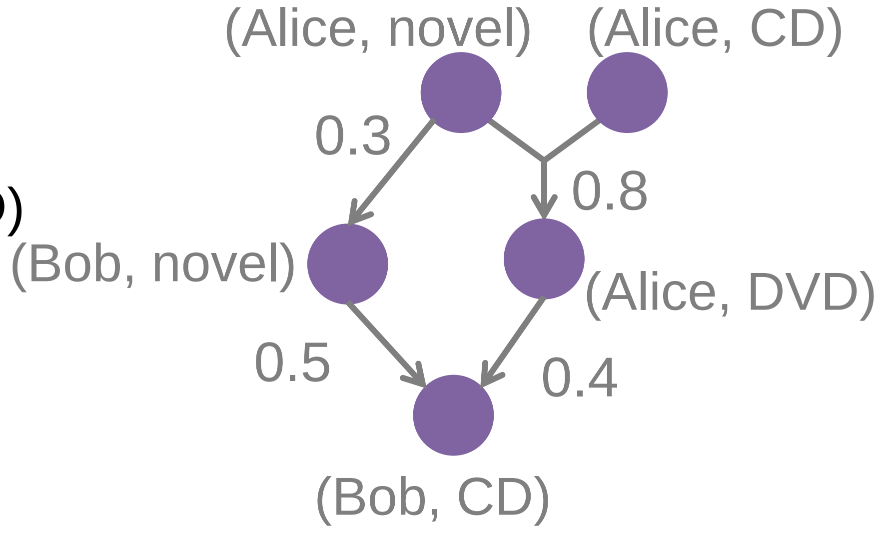
Social Item Graph – A Generalized Model

- For a social network $G(V,E)$ and a set of product items I, the Social Item Graph (SIG) has
  - Node: (v, i) The purchase of a user v for an item i
  - Hyperedge: $X \rightarrow (v', i')$ with probability $p_{v',i',X}$
    - Decision jointly from multiple nodes
    - E.g., 1) Alice buys novel and CD →Alice buys DVD (0.8)
      2) Alice and Bob buy the CD → Cindy buys the CD (0.5)
    - May cross different users/items



Diffusion Process in SIG

- Diffusion Process in SIG
  - Initially, only seeds are activated
  - For a hyperedge $X \rightarrow (v', i')$, if all source nodes in X are activated, then (v', i') has a single chance to be activated with probability $p_{v',i',X}$



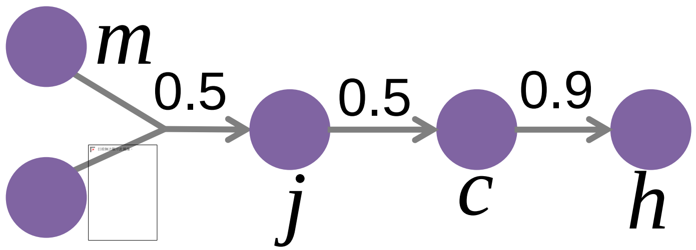
- An example
  - Seed: (Alice,novel), (Alice, CD)
  - Iteration 1:
    - (Bob, novel): 0.3
    - (Alice, DVD): 0.8
  - Iteration 2:
    - (Bob, CD): 1-(1-0.3×0.5)(1-0.8×0.4)=0.422

Social Item Maximization Problem

- Social Item Maximization Problem(SIMP)
  - Given a SIG $G_{SI}\,(V_{SI},E_H)$
  - Select k seeds (nodes in V_{SI})
  - Maximize the number of total activated nodes
- A possible weighted extension of SIMP
  - Assign different profits for each product
  - Maximize the total revenue

Non-submodularity

- SIMP is non-submodular
  - The *1-1/e* approximation ratio does not hold



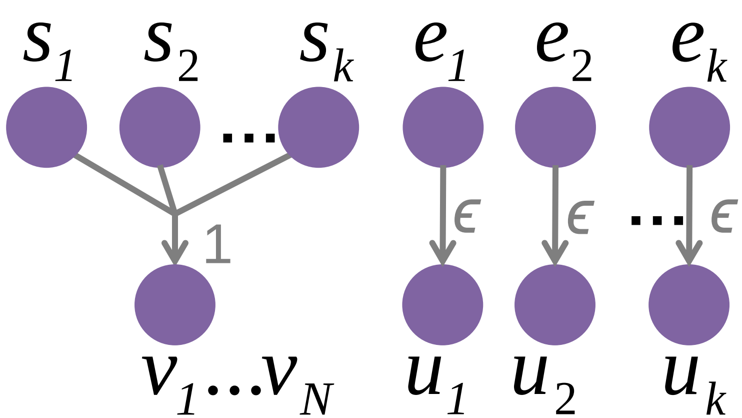
Seed(s)	P(b)	P(m)	P(j)	P(c)	P(h)	Spread
c	0	0	0	1	0.9	1.9
c, b	1	0	0	1	0.9	2.9
c, m	0	1	0	1	0.9	2.9
c, m, b	1	1	0.5	1	0.9	4.4

1

1.5

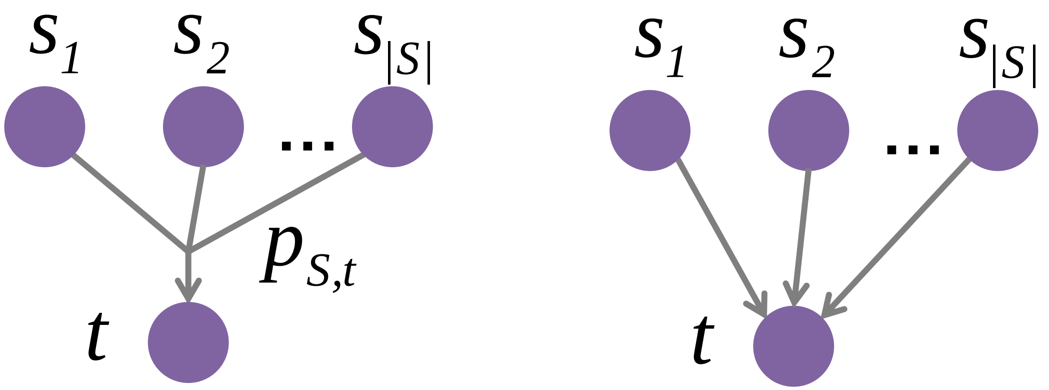
Poor Greedy Performance

- Greedily select a node with the largest increment as a seed may perform poorly
  - The ratio (optimal/greedy) ~ N/k



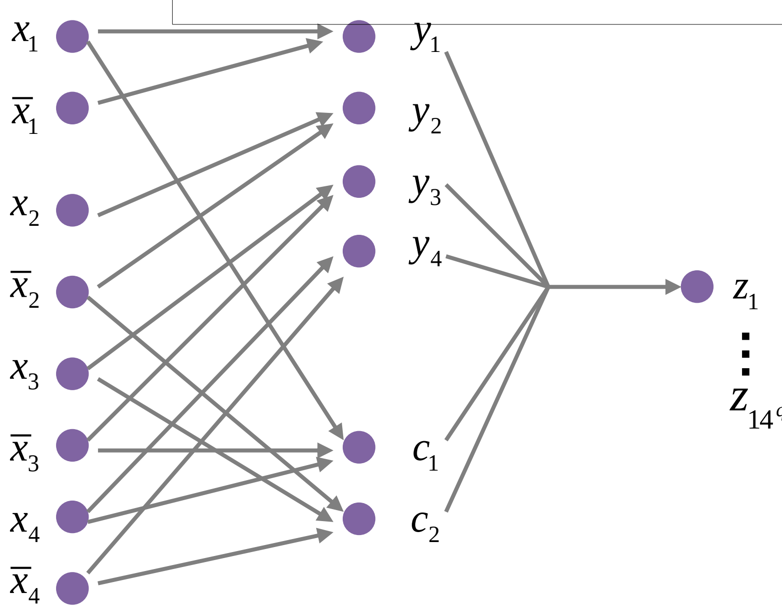
Graph Transformation

- Transforming G_{SI} to a single graph is also a possible solution
  - It cannot handle the case if the activation probability of s_m approaches zero



Hardness Result

- SIMP is inapproximable within n^c for any $c<1$
  - A gap-introducing reduction from 3SAT (NP-complete)
    - Transform an expression ϕ to a SIG G_{SI}
    - If ϕ is satisfiable, $OPT(G_{SI}) \geq (m_{cla}+3n_{var})^q$
    - If ϕ is not satisfiable, $OPT(G_{SI}) < m_{cla}+3n_{var}$
  - An example of 



Hardness Result

- There is no ($(m_{cla}+3n_{var})^{q-1}$)-approximation algorithm
  - Otherwise the 3-SAT can be solved in polynomial time
- For any $\varepsilon>0$, choose $q \geq 2/\varepsilon$, then
- Thus, there exists no ($n^{1-\varepsilon}$)-approximation algorithm

Algorithm Design

- Hyperedge-Aware Greedy (HAG) Strategy
  - Select multiple seeds in a iteration
    - $C(|V_{SI}|, x)$ combinations if selecting x seeds → not feasible
    - Only consider source combinations
      - A source combination includes all sources of a hyperedge
  - Complexity: $O(k \times |E_H| \times c_{dif})$
    - k: # of iterations
    - c_{dif}: the diffusion cost

Algorithm Design (cont.)

- Cross-Edge Selection (CES) Strategy
  - Consider multiple hyperedges jointly
  - For a source combination X, extract a node v
    - With maximal total weight of included hyperedges
  - Consider the combination $X+\{v\}$
  - Time complexity: $O(k \times |E_H| \times |V_{SI}| \times c_{dif})$

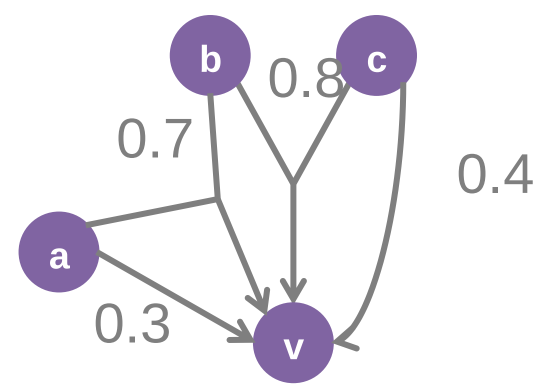
- CES is an n-approximation algorithm of SIMP

Acceleration of Diffusion Computation

- A large hyperedge contains many small ones
  →Exponential number of hyperedges for a node v
  – A new challenge in SIMP
- Pre-process SIG before selecting seeds to accelerate

Definition: The activated probability of v at i_t is

- N_v^a : active neighbors
- N_v^{new} : neighbors activated in i_{t-1}



Acceleration of Diffusion Computation (cont.)

Definition: The aggregated probability of v when all nodes in S are active is

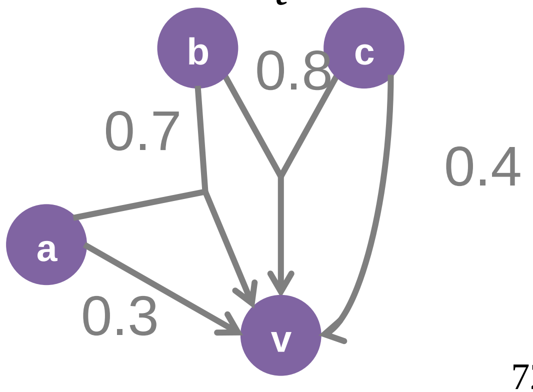
$$\bar{p}_{S,v} = 1 - \prod_{S' \subseteq S, (S' \to v) \in E_H} (1 - p_{S',v})$$

- Pre-compute $\bar{p}_{S,v}$ for any hyperedge $S \to v$

Definition : The activation probability of v at i_t is

$$ap_{v,i_t} = 1 - \frac{1 - \bar{p}_{N_v^a, v}}{1 - \bar{p}_{N_v^{old}, v}}$$

- N_v^a : active neighbors
- N_v^{old} : neighbors active before i_{t-1}

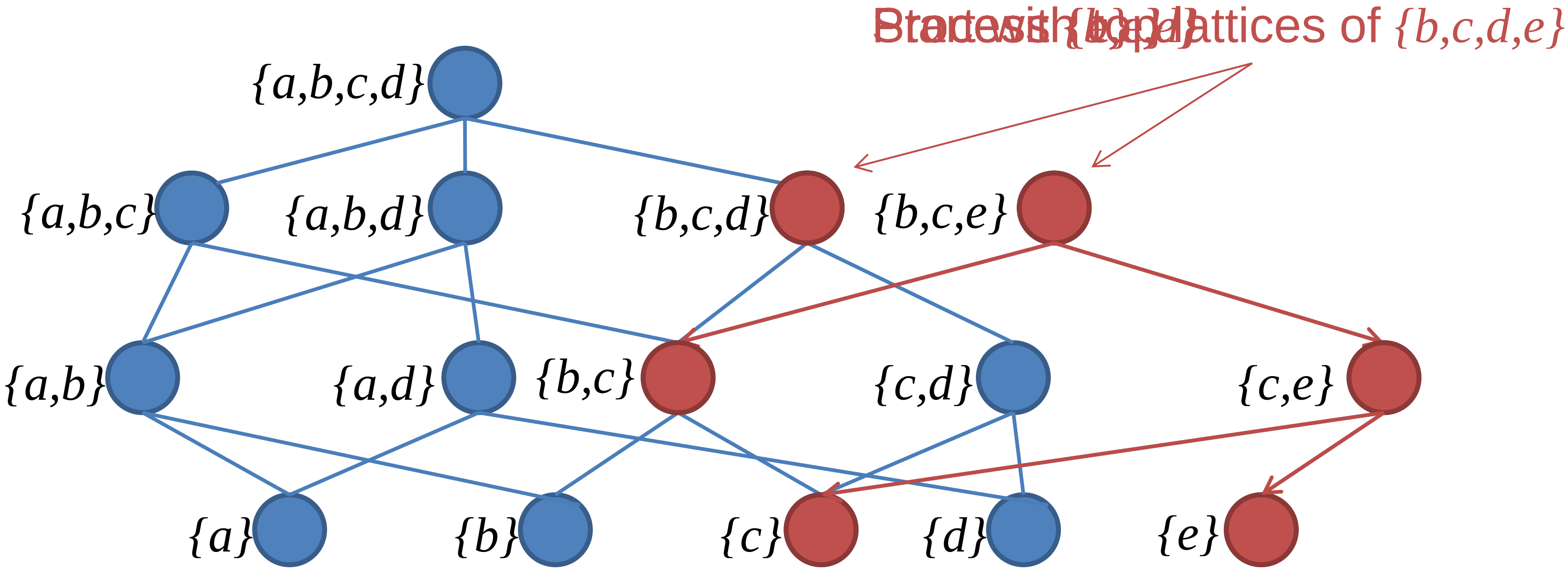


Acceleration of Diffusion Computation (cont.)

- Traversing the Lattice Cache to obtain $\overline{p}_{S,v}$
  - A queue Q for maintain the vertex to be examined
  - Initially, Q contains the top lattice nodes of S
  - Strategy for vertex X
    - If X does not overlap with S, disregard X
    - If X overlaps with S but is not a subset of S, insert the child nodes into Q
    - If X is a subset of S
      - Adopt  , if X overlaps with an examined vertex Y
      - Adopt $p_{X,v}$, otherwise
- Time complexity (c_{dif})
  - Building the lattice: $O(|E_H|^2)$
  - Traversing: $O(|E_H|)$

Acceleration of Diffusion Computation (cont.)

- Derivate the aggregated probability of $S=\{b, c, d, e\}$



Q $\{b,c,d\}$ $\{b,c,e\}$ $\{b,c\}$ $\{c,e\}$ $\{c\}$ $\{e\}$

Adopting  $\{b,c,d\}$

A subset of S without overlapping $\{b,c,d\}$

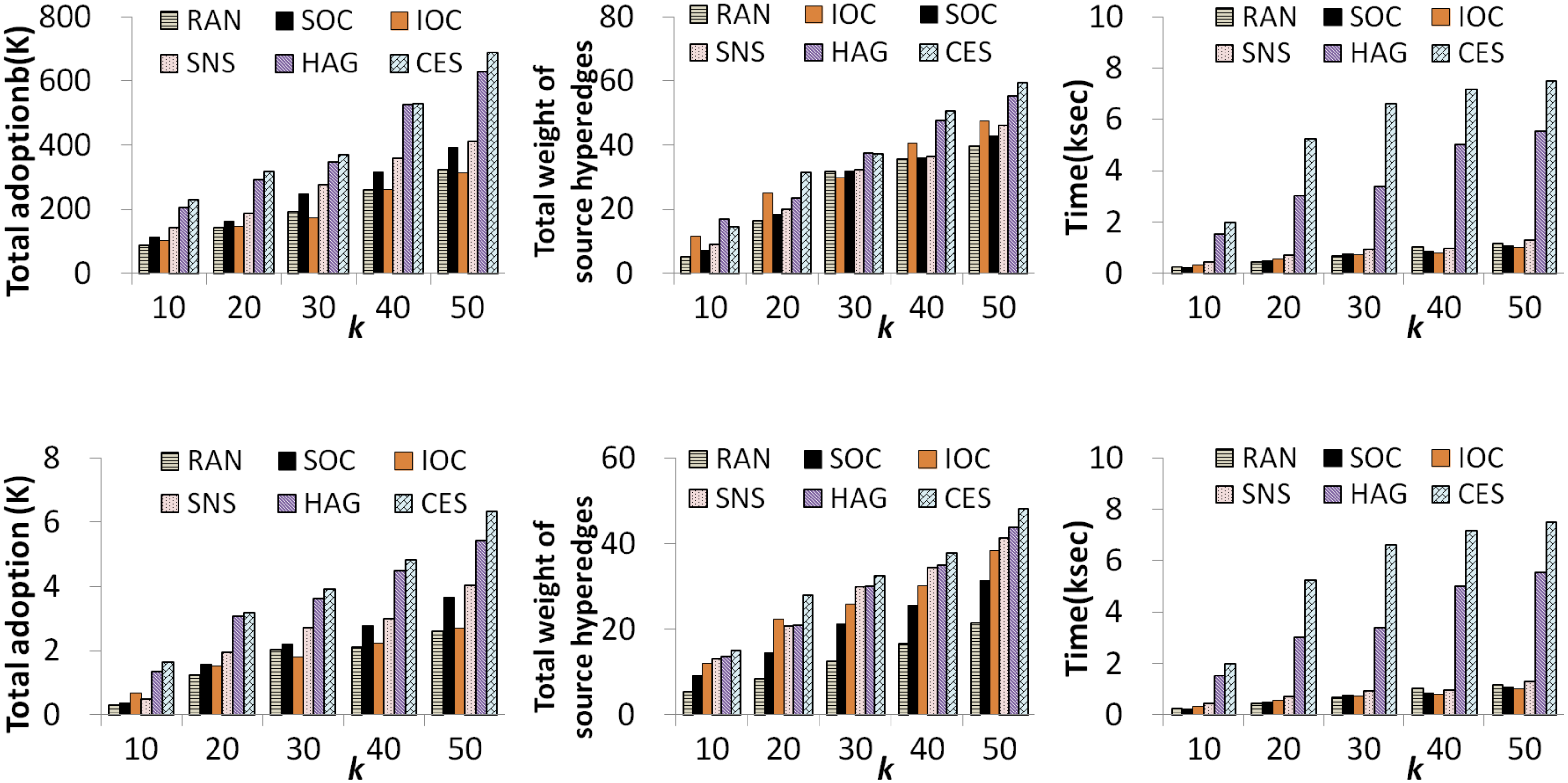
Adopting $p_{X,v}$ $\{b,c,e\}$ $\{c,e\}$ $\{e\}$

Experiments

- Real dataset
  - Douban
    - 5520243 users, 86343003 links, 14050265 bookmarks
  - Foursquare
    - 18107 users, 115574 links, 20734170 check-ins
- Baselines
  - Random approach(RAN)
  - Single node selection approach (SNS)
  - Social approach (SOC)
  - Item approach (IOC)

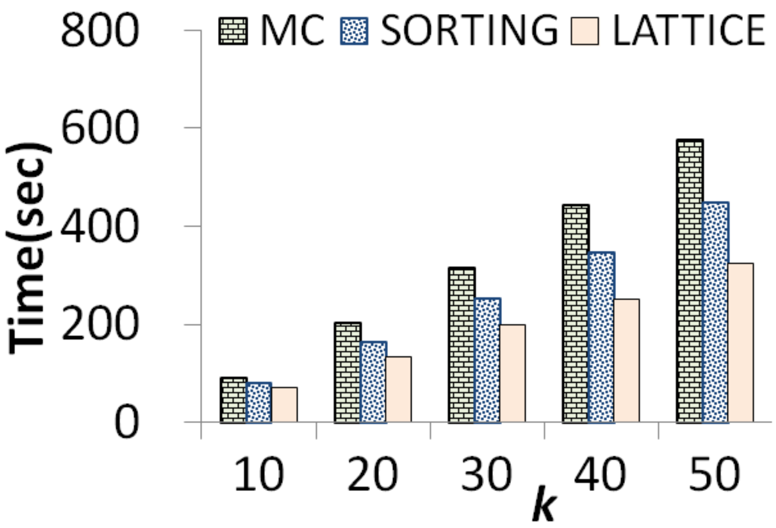
Experiments

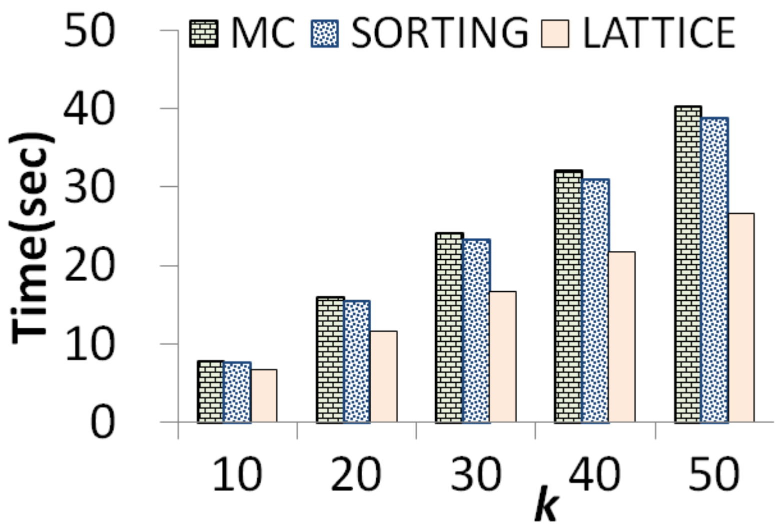
- HAG & CES



Experiments (cont.)

- Lattice Caching





Social Temporal Group Search

- Automatic activity planning service is desirable
  - Tedious manually coordination via email, phone, messenger
  - The time is ripe
- New query: social-temporal group query (STGQ)
  - Given activity size, length, social radius and acquaintance limit
  - Identify a set of activity *attendees* and suitable *time slots*
  - Minimize the total social distance (NP-hard)
- ILP formulation and algorithm design
  - Radius graph extraction, access ordering
  - Pivot time slot, distance pruning, acquaintance pruning

D.-N. Yang, Y.-L. Chen, W.-C. Lee, and M.-S. Chen, "On Social-Temporal Group Query with Acquaintance Constraint," VLDB, 2011.

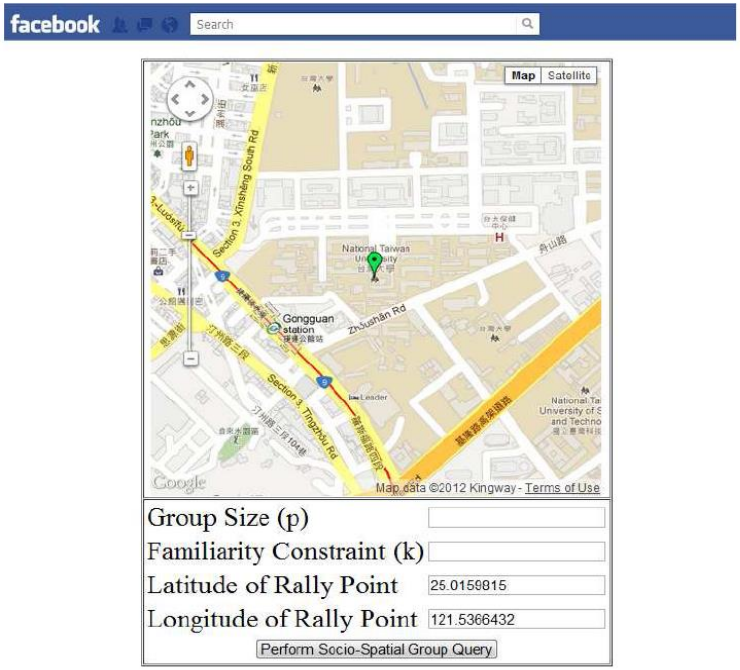
Social Group Search – From Temporal to Spatial

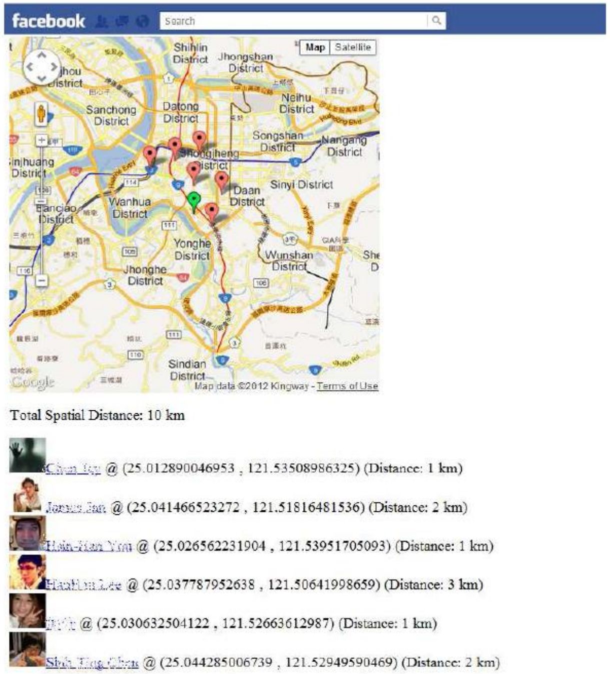
- Extension to impromptu activity organization
  - Locations of nearby friends
- New query: social-spatial group query (SSGQ)
  - Given rally point, activity size, social acquaintance limit
  - Minimize the total spatial distance to the rally point (NP-hard)
- Social R-Tree
  - Hierarchically cache the social and spatial info
  - Organize the social info with different social limits
- ILP formulation and algorithm design

D.-N. Yang, C.-Y. Shen, W.-C. Lee, and M.-S. Chen, "On Socio-Spatial Group Query for Location-Based Social Networks," ACM KDD, 2012.

Implementation and User Study

- User study of 206 people (manually coordination)
  - SSGQ is much faster (mini-seconds v.s. seconds or minutes)
  - SSGQ solutions are better (25%-50% distance reduction)
  - SSGQ solutions are more accurate (100% v.s. 60-80%)





Social Group Search – From Temporal to Spatial

- One rally point to multiple rally points
  - SSGQ: find a group to minimize spatial distance to a rally point
  - MRGQ: find a pair of a group and a location which incurs the minimum spatial distance among all possible pairs
- New query: Multiple rally-point social spatial group query (MRGQ)
  - Given a set of rally points, activity size, social acquaintance limit
  - Find a <group, rally point> pair that has the minimum spatial distance
- Hardness
  - NP-Hard but polynomial-time solvable in Threshold Graph
- Indexing and pruning
  - Indexing users with R-Tree, indexing rally points with BallTree
  - Socio-spatial ordering, All-pair distance ordering, Inner-triangle distance pruning, outer-triangle distance pruning, activity location distance pruning

C.-Y. Shen, D.-N. Yang, L.-H. Huang, W.-C. Lee, and M.-S. Chen, "Socio-Spatial Group Queries for Impromptu Activity Planning," IEEE TKDE, 2015.

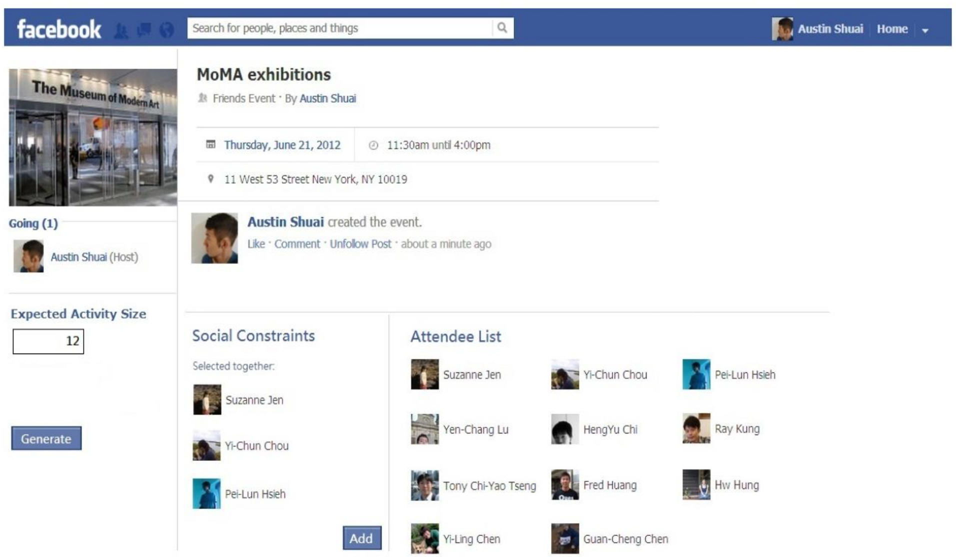
Willingness Optimization for Group Search

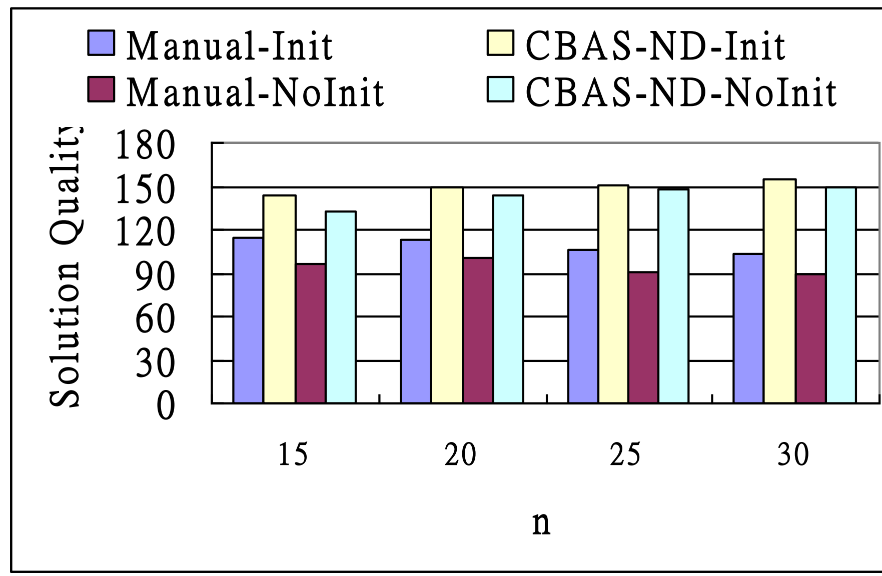
- Willingness optimization
  - Interest + social
  - $$\max_F W(F) = \max_F \sum_{v_i \in F} (\eta_i + \sum_{v_j \in F : e_{i,j} \in E} \tau_{i,j})$$
  - NP-Hard (with reduction from DkS)
- Parameter settings for varied scenarios
  - Friend and foe
  - Exhibition and concert
- Connected or disconnected, which one more difficult?
  - Previous two works return unconnected social groups
  - disconnected -> connected

H.-H. Shuai, D.-N. Yang, P. S. Yu, and M.-S. Chen, "Willingness Optimization for Social Group Activity," VLDB, 2014.

Willingness Optimization for Group Search

- Randomized algorithm with a performance bound
  - Selection of seed nodes
  - Computation budget allocations of seed nodes with sampling
  - Neighbor differentiation with cross-entropy distance





A Comprehensive Study on Willingness Maximization for Social Activity Planning with Quality Guarantee

- Willingness and activity cost optimization
  - Interest + social + activity cost

$$U(F) = \sum_{v_i \in F} (\eta_i + \sum_{v_j \in F : e_{i,j} \in E} \pi_{i,j}) - \beta C(|F|),$$

  - Still NP-hard (with reduction from WASO)
  - The number of enumeration will be 2^n (compared with C_k^n).
  - The user study shows that β is 0.514 on average.
- Randomized algorithm with a performance bound
  - Computational budgets allocation to different sizes and start nodes

H.-H. Shuai, D.-N. Yang, P. S. Yu, and M.-S. Chen, " A Comprehensive Study on Willingness Maximization for Social Activity Planning with Quality Guarantee," IEEE TKDE, 2015.

Spatial-Proximity Optimization for Rapid Task Group Search

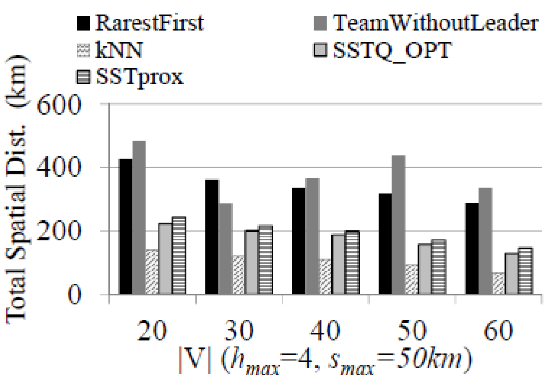
- Crucial factors for task groups
  - Team transport and rapid response (spatial domain)
  - Team member skills (skill domain)
  - Team social rapport (social domain)
- New Query: Spatio-Social Team Query (SSTQ)
  - Given required skill set, query point, hop constraint, spatial constraint
  - Find a group while covering the required skills, satisfying hop and spati
    constraints, and minimizing the total spatial distance to the querying poi
- Problem Analysis
  - NP-Hard
  - Inapproximable within any factor unless P=NP

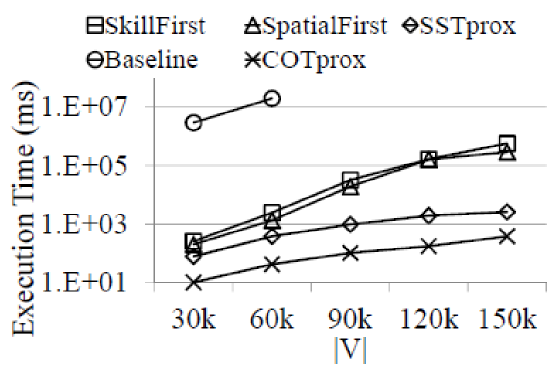
C.-Y. Shen, D.-N. Yang, W.-C. Lee, and M.-S. Chen, "Spatial-Proximity Optimization for Rapid Task Group Deployment," ACM TKDD, 2015.

Spatial-Proximity Optimization for Rapid Task Group Search

- Proposed algorithms
  - A ln|*T*| approximation algorithm with guaranteed error bound (SSTprox)
    - T is the set of required skills
  - Two database query algorithms that finds the optimal solutions (SkillFirst, SpatialFirst)
    - With effective ordering and pruning strategies

- Performance Evaluation





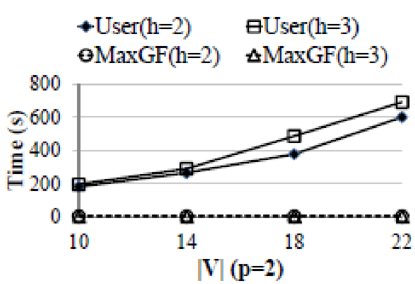
Maximizing Friend-Making Likelihood for Socialization Group Search

- For organizing socialization activities
  - Face-to-face friend-making (dating) activiti
  - Via online social network services
- Gap between existing activity organization and friend recommendation in OSNs
  - <u>Activity organization</u>: extracting socially cohesive groups from OSNs
  - <u>Friend recommendation</u>: finding potential new friends
- Model the social network as heterogeneous graph
  - Individuals (vertex), existing friend (friend edge), potential friend (potential edge)
  - Weights on potential edges: friend-making likelihood (obtained from link prediction algorithms)

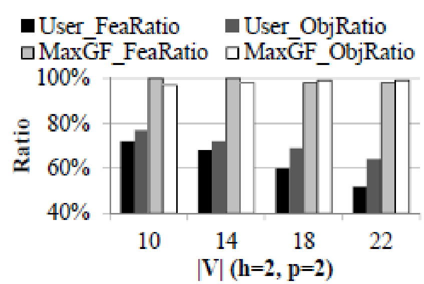
C.-Y. Shen, D.-N. Yang, W.-C. Lee, and M.-S. Chen, "Maximizing Friend-Making Likelihood for Social Activity Organization," PAKDD, 2015. (<u>Best Paper Runner-Up Award</u>)

Maximizing Friend-Making Likelihood for Socialization Group Search

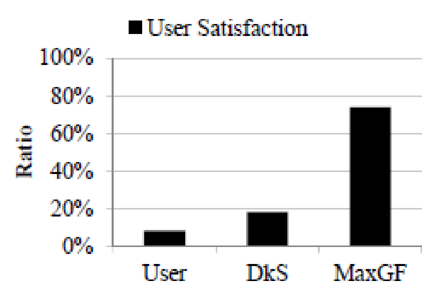
- New problem: hop-bounded maximum group friending
  - Given heterogeneous social graph, hop and group size constraints
  - Find a group that maximizes the total weight on incident potential edge while ensuring the social tightness (hop constraint), and the group is sufficiently large
  - NP-Hard and inapproximable within any factor
- Algorithm design
  - 3-approximation algorithm with guaranteed error bound
- User study



(a) Required Time.



(b) FeaRatio and ObjRatio.



(c) User Satisfaction.

Viral Marketing – Seed Search

- Viral marketing in social networks
  - Word-of-mouth social influence via social network applications
  - Previous: spread maximization for a single product
    - Seed selection problem for broadcasting (only social dimension)
- Our observations
  - Product purchase decisic
    - Social dimension + *preference dimension*
  - Product *bundling*
- New problem: product bundling
  - Choosing *a given number of product items* for spread maximization
  - NP-Hard (from frequent patterns mining)

livingsocial
大众点评 dianping.com
新浪微博 weibo.com
renren
facebook
KLOUT

D.-N. Yang, W.-C. Lee, N.-H. Chia, M. Ye and H.-J. Hung, "On Bundle Configuration for Viral Marketing in Social Networks," ACM CIKM, 2012.

Active Friending – Intermediate Search

- No active friending service exists in social networking websites
  - Existing websites suggest possible friends passively
  - A user may want to make friend with a desired one actively
- New problem: acceptance probability maximization (APM)
  - Given <u>initiator</u>, <u>friending target</u>, and <u>invitation budget</u>
  - Identify a set of users to send invitations iteratively
  - Maximize the acceptance probability of the target
  - NP-hard (not in APX) in general graphs
- Algorithm design
  - A polynomial-time algorithm to find the optimal solution in MIA

D.-N. Yang, H.-J. Hung, W.-C. Lee, and W. Chen, "Maximizing Acceptance Probability for Active Friending in On-Line Social Networks," ACM KDD, 2013. (<u>featured by MIT Technology Review and ACM TechNews</u>)

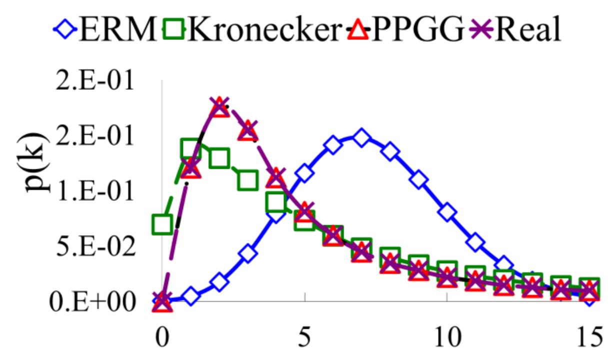
Pattern-Preserving Social Graph Generator

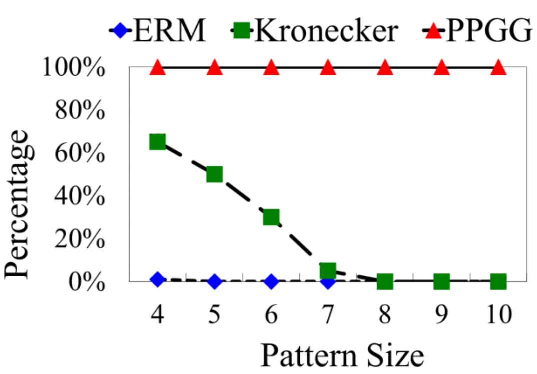
- Most social real datasets contain only millions of nodes
  - Difficulty in crawling real datasets in social websites
- The problem of generating synthetic graphs is to maintain the distinguishing characteristics of real-world networks
  - Node degree, degree distribution, diameter, and clustering coefficient
- However, no one has aimed to preserve the *frequent patterns* in data mining for synthetic graphs

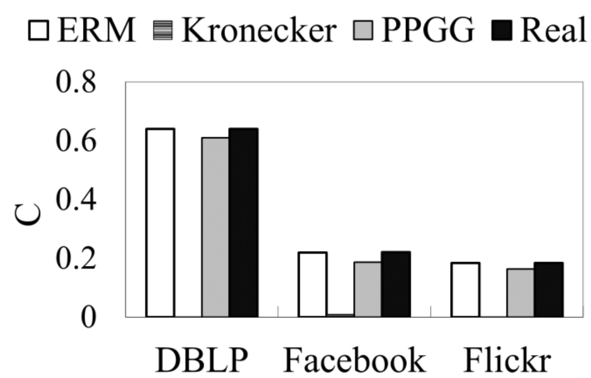
H.-H. Shuai, D.-N. Yang, P. S. Yu, C. -Y. Shen, and M.-S. Chen, "On Pattern Preserving Graph Generation," IEEE ICDM, 2013.

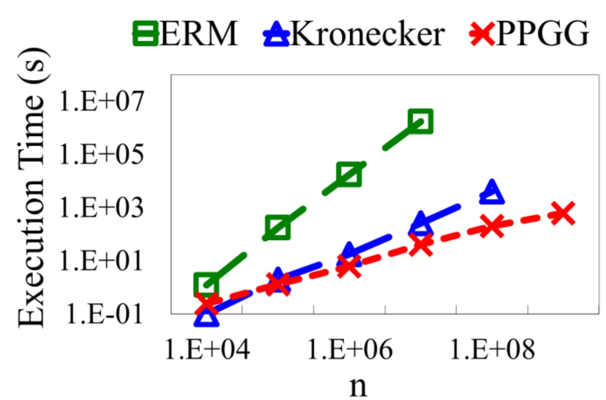
Pattern-Preserving Social Graph Generator

- We propose a *Pattern Preserving Graph Generator (PPGG)*
  - Large single unlabeled graph with the target node number, degree distribution, and clustering coefficient, and the frequent patterns with the required supports
- PBGG contains two phases:
  - Phase 1: Pattern Overlapping Phase
  - Phase 2: Graph Augmentation Phase
- Generate a billion-node graph in mins









Detecting Social Network Mental Disorders and Forming Therapy Groups

- Social Network Mental Disorders (SNMDs)
  - Cyber-Relationship Addiction, Information Overload, Net Compulsion
  - Usually observed passively (e.g., by teachers or parents)

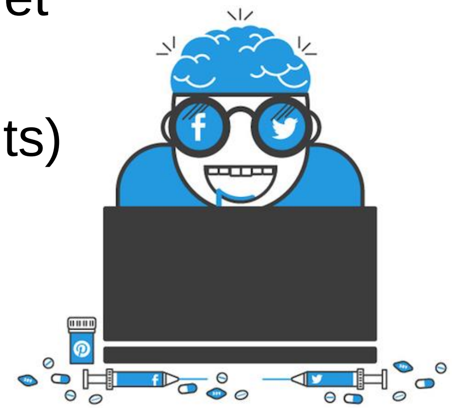


Image source: http://www.mediabistro.com/alltwitter/tag/social-media-addiction

- To detect such mental disorders in Early Stage
  - Online usage time only moderately correlated
  - Propose SNMD Detection (SNMDD) framework
- And to Form therapy groups for the identified patients
  - Three important criteria (i) unfamiliarity of patients, (ii) similarity of symptoms, (iii) therapy group size
  - Formulate Patient Selection for Group Therapy (PSGT) problem

C.-Y. Shen, H.-H. Shuai, D.-N Yang, Y.-F Lan, W.-C. Lee, P. S. Yu, and M.-S. Chen, ``Forming Online Support Groups for Internet and Behavior Related Addictions ,'' ACM CIKM 2015.

Detecting Social Network Mental Disorders and Forming Therapy Groups

- Features extracted for detecting SNMDs
  - **Social Interaction Features:** Parasocial relationship, On/Off-line ratio, Social Capital, Social Searching and Browsing,
  - **Personal Features:** Self-Disclosure Based Features (Selfies, Emoticons, Stickers, Ratio between Like and Comment), Temporal Behavior Features, Disinhibition Based Features, Profile Features
  - Employ SVM classifier for prediction
- Patient Selection for Group Therapy (PSGT)
  - Given social network and the similarities among each pair of patients
  - Find a subgraph ***H*** such that: **1)** each pair of patients in ***H*** are neither friends, nor friend of friend; **2)** H has no fewer than ***p*** nodes; **3)** maximize the similarity of the selected patients in ***H***
  - PSGT in NP-Hard to solve, and inapproximable within any factor
  - Propose an error-bounded 3-approximation algorithm

Reference

- D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 1998.
- A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 1999.
- J. Kleinberg. Navigation in a small world. *Nature*, 2000.
- R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 2012.
- D. Kempe, J. Kleinberg, E. Tardos. Maximizing the spread of influence through a social network. In KDD, 2003.
- S. Bhagat, A. Goyal, L. V. S. Lakshmanan. Maximizing product adoption in social networks. In WSDM, 2012.
- 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.

Reference (cont.)

- X. Hey, G. Song, W. Chen and Q. Jiang. Influence blocking maximization in social networks under the competitive linear threshold model. In SDM, 2012.
- M. Eftekhar, Y. Ganjali, and N. Koudas. Information cascade at group scale. In KDD, 2013.
- J. Tang, S. Wu, and J. Sun. Confluence: conformity influence in large social networks. In KDD, 2013.
- A. Goyal, F. Bonchi. L. V. S. Lakshmanan. Learning influence probabilities in social networks. In WSDM, 2010.
- W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In KDD, 2010.
- W. Chen, Y. Yuan, and L. Zhang. Scalable influence maximization in social networks under the linear threshold model. In ICDM, 2010.

Reference (cont.)

- D.-N. Yang, Y.-L. Chen, W.-C. Lee, and M.-S. Chen, "On Social-Temporal Group Query with Acquaintance Constraint," VLDB, 2011.
- C.-H. Tai, P. S. Yu, D.-N. Yang, and M.-S. Chen, "Privacy-Preserving Social Network Publication against Friendship Attacks," ACM KDD, 2011.
- D.-N. Yang, C.-Y. Shen, W.-C. Lee, and M.-S. Chen, "On Socio-Spatial Group Query for Location-Based Social Networks," ACM KDD, 2012.
- D.-N. Yang, W.-C. Lee, N.-H. Chia, M. Ye and H.-J. Hung, "On Bundle Configuration for Viral Marketing in Social Networks," ACM CIKM, 2012.
- D.-N. Yang, H.-J. Hung, W.-C. Lee, and W. Chen, "Maximizing Acceptance Probability for Active Friending in On-Line Social Networks," ACM KDD, 2013. (featured by MIT Technology Review and ACM TechNews)
- G. Ference, W.-C. Lee, H.-J. Hung, and D.-N. Yang, "Spatial Search for K Diverse-Near Neighbors,“ ACM CIKM, 2013.
- H.-H. Shuai, D.-N. Yang, P. S. Yu, C. -Y. Shen, and M.-S. Chen, "On Pattern Preserving Graph Generation,“ IEEE ICDM, 2013.
- X. Liu, D.-N. Yang, M. Ye, and W.-C. Lee, "U-Skyline: A New Skyline Query for Uncertain Databases," IEEE TKDE, 2013.

Reference (cont.)

- H.-H. Shuai, D.-N. Yang, P. S. Yu, and M.-S. Chen, "Willingness Optimization for Social Group Activity," VLDB, 2014.
- C.-H. Tai, P. S. Yu, D.-N. Yang, and M.-S. Chen, "Structural Diversity for Resisting Community Identification in Published Social Networks," IEEE TKDE, 2014.
- C.-Y. Shen, D.-N. Yang, W.-C. Lee, and M.-S. Chen, "Maximizing Friend-Making Likelihood for Social Activity Organization," PAKDD, 2015. (Best Paper Runner-Up Award)
- C.-Y. Shen, D.-N. Yang, L.-H. Huang, W.-C. Lee, and M.-S. Chen, "Socio-Spatial Group Queries for Impromptu Activity Planning," IEEE TKDE, 2015.
- C.-Y. Shen, D.-N. Yang, W.-C. Lee, and M.-S. Chen, “Spatial-Proximity Optimization for Rapid Task Group Deployment,” ACM TKDD, 2015.
- H.-H. Shuai, D.-N. Yang, P. S. Yu, and M.-S. Chen, " A Comprehensive Study on Willingness Maximization for Social Activity Planning with Quality Guarantee,” IEEE TKDE, 2015.
- C.-Y. Shen, H.-H. Shuai, D.-N Yang, Y.-F Lan, W.-C. Lee, P. S. Yu, and M.-S. Chen, ``Forming Online Support Groups for Internet and Behavior Related Addictions ," ACM CIKM 2015