Enabling and Controlling Diffusion Processes in Networks

Zhifeng Sun

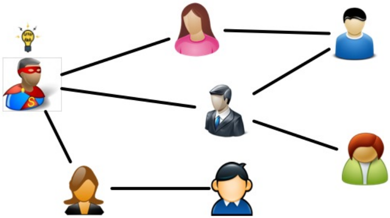
Department of Computer Science
Northeastern University

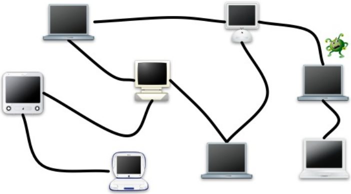
July 19, 2011

Diffusion process

- Diffusion is the spread of information or commodities in the network through local transmissions.
- Harmful/Negative diffusions:
  - Diffuse harmful information (e.g. diseases, viruses).
  - Analyze the converging time and the extend of diffusion processes
  - Design good intervention strategies.
- Positive diffusions:
  - Diffuse useful information (e.g. innovations, ideas).
  - Analyze the converging time of diffusion processes.
  - Design efficient algorithms for fast diffusion.

Motivation





- Innovations, ideas, gossip.
- Diseases.
- Friendship.

- Resource discovery.
- Computer viruses.
- Also sensor networks, mobile networks, etc.

Thesis concentration

- What is the optimal intervention strategy for a given contact network?
- How effective are interventions of individual choices and behaviors.
  - Individuals make their own intervention strategies.
  - Individuals exhibit risk behavior changes.
- Analyze positive diffusions on dynamic networks.
  - Resource discovery in the networks of gossip.
  - Information dissemination in adversarial networks.

Outline

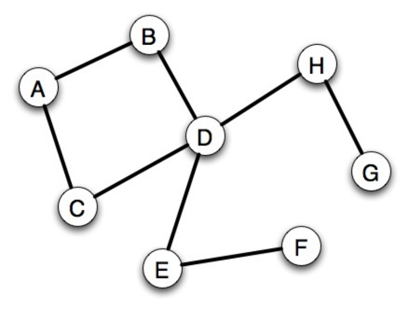
Outline

Model

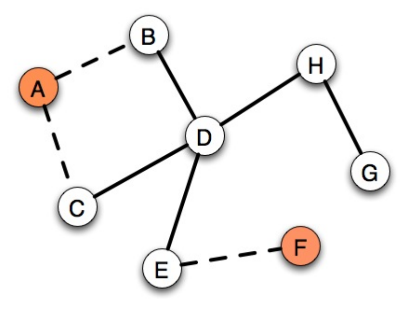
- Contact graph: $G = (V, E)$.
- Intervention: $a_i \in \{0, 1\}$
  - $a_i = 1$: node i takes intervention.
  - $a_i = 0$: node i dosen't take intervention.
- Intervention vector: $\bar{a} = (a_1, a_2, \ldots, a_n)$.
- Intervention cost and infection cost: C_i, L_i.
- Individual cost: $\text{cost}\,(\bar{a}) = a_i C_i + (1 - a_i) p_i(\bar{a}) L_i$, where $p_i(\bar{a})$ is the probability that node i gets infected given $\bar{a}$.
- Social cost: $\sum_i \text{cost}\,(\bar{a})$.
- We assume the infection is initialized at a node randomly picked according to an arbitrary probability distribution $\bar{w} = (w_1, w_2, \ldots, w_n)$.
- Disease transmission locality parameter d: how far the disease can transmit from the source node.

Example ($d = 2$)

Original graph

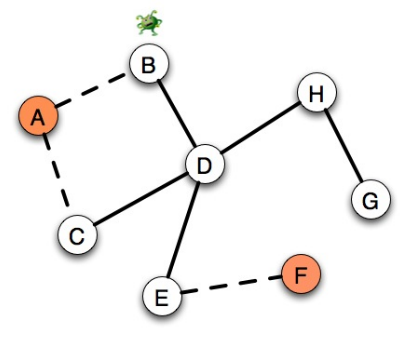


A and F take interventions

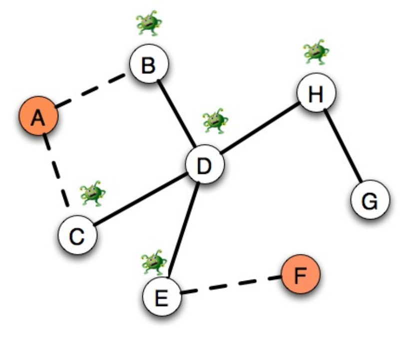


Example ($d = 2$)

B started infection



Spread distance d



Outline

Centralized strategies

Problem definition:

- Given any contact graph, find intervention vector $\bar{a}$ to minimize the social cost $\sum_i \text{cost}(\bar{a}) = \sum_i [a_i C_i + (1 - a_i) p_i(\bar{a}) L_i]$.

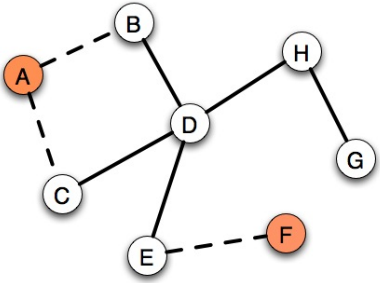
Our results:

- Computing the social optimum is NP-complete for all d.
- Give an LP based approximation algorithm.
  - $d < \infty$: $2d$-approximation.
  - $d = \infty$: $O(\log n)$-approximation.
- Results published in [Kumar-Rajaraman-Sun-Sundaram 2010].

Related work

- [Aspnes-Chang-Yampolsky 2006] introduced a basic model for $d = \infty$ case with uniform intervention and infection costs which we have generalized here.
  - Give an $O(\log^{1.5} n)$-approximation for social optimum.
- [Chen-David-Kempe 2010] independently gave an $O(\log n)$-approximation algorithm.
- [Dezsö-Barabási 2002] studied how to control virus transmission on scale-free networks.
- [Borgs-Chayes-Ganesh 2010] studied how to distribute antidotes to control epidemics.
- Considerable work in SIR and SIS models in epidemiology.

Example for calculating $p_i(\bar{a})$



- Initial infection probability is $1/8$ for all nodes.
- $d = 2$: $p_B(\bar{a}) = 5/8$, and $p_G(\bar{a}) = 3/8$.
- $d = \infty$: $p_B(\bar{a}) = p_G(\bar{a}) = 6/8$.

Approximation algorithm for social optimum

LP formulation

- Let P_{ij}^d denote the set of all simple paths from i to j of length at most d.
- $\forall v \in V$, $x_v = 1$ if v is secure; $x_v = 0$ otherwise.
- $\forall i, j \in V$, $y_{ij} = 1$ is there is no $p \in P_{ij}^d$ consisting entirely of insecure nodes.

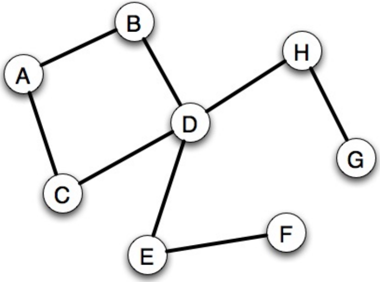
$$\begin{array}{ll} \min & \sum_v C_v \cdot x_v + \sum_{j \in V} L_j \sum_{i \in V} w_i (1 - y_{ij}) \\ \text{s.t.} & \sum_{v \in p} x_v \geq y_{ij} \;\; \forall p \in P_{ij}^d \\ & x_v \in \{0, 1\} \; \forall v \in V \\ & y_{ij} \in \{0, 1\} \; \forall i, j \in V \end{array}$$

LP in details

$$\begin{array}{ll} \min & \sum_v C_v \cdot x_v + \sum_{j\in V} L_j \sum_{i\in V} w_i(1-y_{ij}) \\ \text{s.t.} & \sum_{v\in p} x_v \geq y_{ij} \;\; \forall p \in P_{ij}^d \end{array}$$

- First part of the objective function corresponds to the cost of securing nodes.
- Second part corresponds to the infection cost. For node j, its infection cost is L_j times the sum of the probabilities of all nodes that have a path to j of length at most d consisting entirely of insecure nodes.
- Constraint says, in order to separate a pair of nodes i and j, we need to secure at least one node in every path between these two.

Example of LP constraints



- $d = 2$.
- Look at the constraints for A, D pair.
  - $x_A + x_B + x_D \geq y_{AD}$
  - $x_A + x_C + x_D \geq y_{AD}$

Algorithm overview

- Solve the LP, and obtain fractional solutions (x, y).
    - d is a constant:
        - Number of paths of length at most d is polynomial.
    - d is not a constant:
        - Number of paths superpolynomial; still LP solvable using ellipsoid method.
- Partial rounding to obtain integral y values.
- Final rounding to obtain integral x values.
- Show the cost of integral solution is within $2d$ or $O(\log n)$ factor of the optimal LP solution.

Partial rounding

$$\begin{array}{ll} \min & \sum_v C_v \cdot x_v + \sum_{j \in V} L_j \sum_{i \in V} w_i (1 - y_{ij}) \\ \text{s.t.} & \sum_{v \in p} x_v \geq y_{ij} \quad \forall p \in P_{ij}^d \end{array}$$

- Let (x, y) denote an optimal solution.
- Round each y_{ij} to nearest integer.
  - So values at least 1/2 are rounded up to 1 and less than 1/2 rounded down to 0.
- Scale up each x_v by a factor of 2.
  - If scaled value exceeds 1, set it to 1.
- New solution (x, y) is still feasible and new cost at most twice that before.

Final rounding

$$\sum_{v \in p} x_v \geq y_{ij} \quad p \in P_{ij}^d$$

- It remains to round the x-values.
- Simple approach: Each x_v that is at least $1/d$ is rounded up to 1, other x_vs rounded down to 0.
  - Yields $2d$-approximation.
  - Perhaps acceptable for small d.
- For $d = \infty$:
  - Need to select a set of nodes to secure such that all pairs of nodes i, j with $y_{ij} = 1$ are separated.
  - This is precisely a vertex multicut problem for which x-values give a fractional optimum.
  - Use algorithm of Garg-Vazirani-Yannakakis to round the x-values and obtain an $O(\log n)$-approximation.

Outline

Models:

- We use game theoretic analysis.
- Strategy for each node is either taking the intervention ($a_i = 1$) or not ($a_i = 0$).
- Utility for each node is the cost function $\text{cost}\,(\bar{a}) = a_i C_i + (1 - a_i) p_i(\bar{a}) L_i$

Our results (published in [Kumar et al 2010]):

	$d = 1$	$1 < d < \infty$	$d = \infty$
existence of pure NE	Yes	No/NP-complete	Yes
price of anarchy	$\Delta + 1$		$O(1/\alpha(G))$

- Δ is the max degree in the contact graph.
- $\alpha(G)$ is the vertex expansion of the contact graph.

Related work

- [Aspnes et al 2006] introduced a basic model for $d = \infty$ case that we have generalized here.
  - Show existence of pure NE in a uniform version.
- [Kearns-Ortiz 2004] introduced interdependent security games.
  - Similar to our model for special case of $d = 1$.
- [Bauch-Earn 2004] used game theory to analyze vaccination uptake level to eradicate diseases.
- [Omic et al 2009] introduced *n*-intertwined games.
  - Based on SIS model for worm spread.
- [Grossklags-Christin-Chuang 2008] introduced information security games.

Existence of pure NE when $d = \infty$

Theorem

There is a pure NE when $d = \infty$.

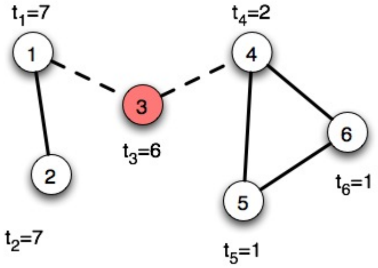
- The existence proof is a potential function argument.
- Define Threshold of a node, t_i: Bound on number of reachable nodes that would make the node want to secure itself.

$$C_i \text{ vs } L_i(t_i + 1)/n \implies t_i = nC_i/L_i - 1$$

- w.l.o.g., assume $t_1 \geq t_2 \geq \cdots \geq t_m$.

Potential function

- Define potential function: $\hat{\Phi}(\vec{a}) = (\Phi_1(\vec{a}), \Phi_2(\vec{a}), \ldots, \Phi_n(\vec{a}))$ where $\Phi_i(\vec{a})$ is 0 if i is secure, -1 if i is insecure and happy, and 1 otherwise.



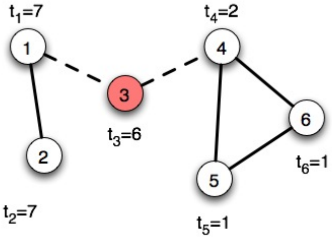
- $t_1 = 7, t_2 = 7, t_3 = 6, t_4 = 2, t_5 = 1, t_6 = 1$.
- 3 is secured.
- Potential function for this configuration is $(-1, -1, 0, -1, 1, 1)$.

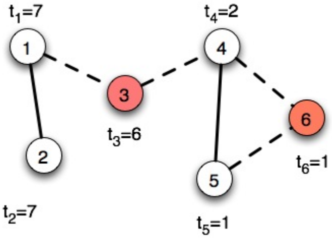
Proof overview

- Start with an arbitrary strategy vector $\bar{a}$.
- Show potential function $\hat{\Phi}(\vec{a})$ decreases lexicographically when everyone does best response.
- There is a lower bound on the potential function, thus will reach a stable value.
- Everyone is satisfied with current strategy (pure NE).

$\hat{\Phi}(\vec{a})$ lexicographically decreases

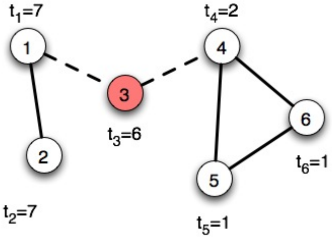
- Case 1: unhappy insecure → happy secure. One component decreases by 1, while none of the other components increases.





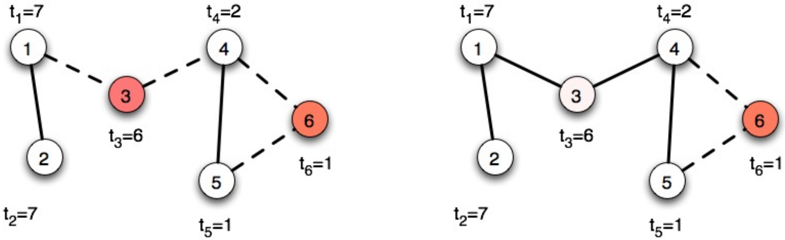
$\hat{\Phi}(\vec{a})$ lexicographically decreases

- $\hat{\Phi}(\vec{a}) : (-1, -1, 0, -1, 1, 1) \rightarrow (-1, -1, 0, -1, -1, 0)$



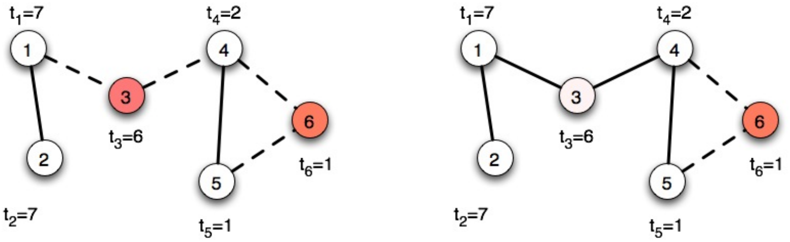
$\hat{\Phi}(\vec{a})$ lexicographically decreases

- Case 2: unhappy secure $\rightarrow$ happy insecure. All the happy insecure nodes with bigger thresholds are still happy. Happy insecure nodes with smaller thresholds may become unhappy. But the function still decreases lexicographically.



$\hat{\Phi}(\vec{a})$ lexicographically decreases

- $\hat{\Phi}(\vec{a}) : (-1, -1, 0, -1, -1, 0) \rightarrow (-1, -1, -1, 1, 1, 0)$



Outline

Motivation:

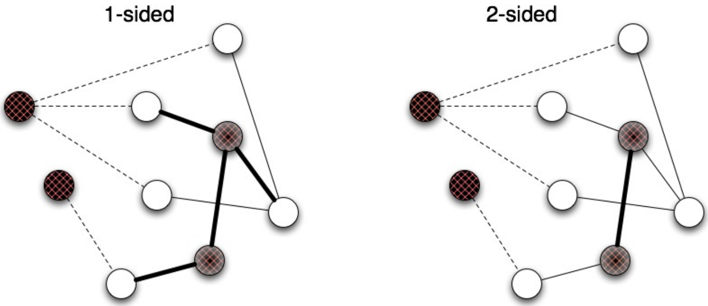
- Drive faster with seat belt on.
- Have more contact when vaccinated.
- Take more risk with government bailout.
- How risk behavior is going to affect intervention strategies?

Model:

- Contact graph $G = (V, E)$.
- Each node either applies intervention or not.
- Intervention succeeds with probability p_s.
    - If succeeds, the node is immune.
    - If fails, the node is still susceptible.
- Disease transmission probability p.

Risk behavior change models

- 1-sided: disease transmission probability on (u, v) is p_m if either u or v is intervention failed node.
- 2-sided: disease transmission probability on (u, v) is p_m if both u and v are intervention failed nodes.

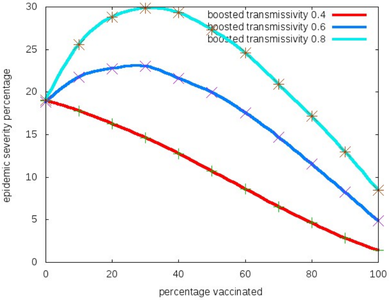


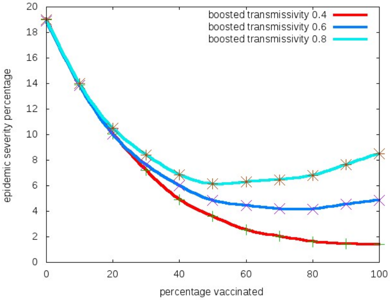
Epidemic size calculation

- Discrete time SIR (susceptible-infected-recovered) model.
- An infected node is assume to recover in one unit of time.
- Each infected node infects its neighbors independently with probability p or p_m.
- Epidemic size is the number of nodes that ever get infected.

Less is more and non-monotonicity

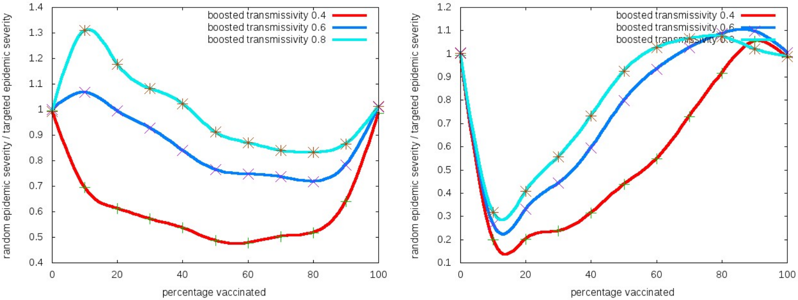
- For both 1-sided and 2-sided risk behavior models, less interventions may be more effective.
- True for both randomized and targeted strategies.
- Simulated on scale-free graphs and Erdös-Rényi random graphs.





Random “may be” better than targeted

- Intervention strategies:
  - Apply interventions to each node uniformly at random.
  - Apply interventions to nodes with high degrees.
- In both 1-sided and 2-sided models, random intervention strategy can be better than targeted strategy.



Ongoing research

- Have rigorous proofs for “less is more” and “random better than targeted” observations.
- Have rigorous proofs on special families of graphs (e.g. Erdös-Rényi random graphs, locally-finite infinite graphs).
- Run simulations on real data sets.

Outline

Resource discovery

- In peer-to-peer networks, nodes can only communicate with those whose IP addresses are known.
- Design efficient distributed algorithm to discover IP addresses on the network.
- The network is altered dynamically by the diffusion process itself.
- Also applies to friendship discovery in social networks.

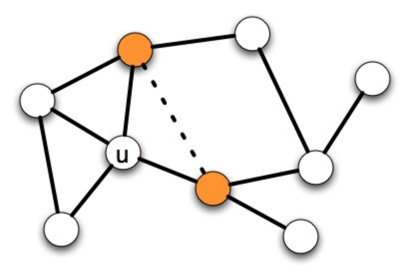
Related work

- [Harchol-balter et al 1999] studied this process with message size $\Omega(n)$, and showed an $O(\log^2 n)$ bound.
- [Law-Siu 2000] gave an $O(\log n)$ randomized algorithm for resource discovery where the message size is $\Omega(n)$.
- [Kutten-Peleg-Vishkin 2003] proposed a deterministic algorithm which solves resource discovery in $O(\log n)$ time but the message size is still $\Omega(n)$.
- [Kutten-Peleg 2002] and [Abraham-Dolev 2006] studied asynchronous resource discovery.

Our algorithms

- Push discovery (triangulation): In each round, each node chooses two random neighbors and connects them by "pushing" their mutual information to each other.
- Notice the message size here is $O(\log n)$.

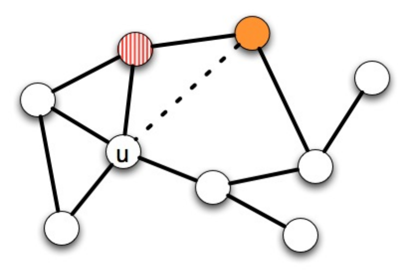
Triangulation process



Our algorithms

- Pull discovery (two-hop walk): In each round, each node connects itself to a random neighbor of one of its randomly chosen neighbors, by "pulling" a random neighboring ID from a random neighbor.
- Notice the message size here is $O(\log n)$.

Two-hop walk process



Ongoing research

- We are interested in the converging time.
- In undirected graphs, we showed the upper bound for both triangulation process and two-hop walk process is $O(n \log^2 n)$, while $\Omega(n \log n)$ is the lower bound.
- In directed graphs, we showed the upper bound for two-hop walk process is $O(n^2 \log n)$, while the lower bound is $\Omega(n^2 \log n)$ for weakly connected graphs and $\Omega(n^2)$ for strongly connected graphs.
- We conjecture that both processes complete in $O(n \log n)$ time in undirected graphs.

Information dissemination in adversarial networks

- k different pieces of information assigned to a set of nodes.
- Goal is to diffuse all k pieces of information to every node on the network.
- We consider adversarial network.

Related work

- [Kuhn et al 2010] studied information dissemination problem in adversarial networks, and showed a tight bound $O(kn)$ in the “shout-out” model with message size $O(\log n)$.
- [Haeupler-Karger 2011] studied the same problem using network encoding.
- [Karp-Schindelhauer-Shenker-Vöcking 2000] introduced pull and push models.
- [Boyd-Ghosh-Prabhakar-Shah 2006] studied randomized gossip algorithms.
- [Mosk-Aoyama-Shah 2006] studied how to compute separable functions via gossip.

Proposed research

- Design efficient algorithms for information dissemination problems in other models.
- Randomized vs deterministic.
- Centralized vs distributed.
- Broadcast vs unicast.
- Resilience of the communication links.
- Power of the adversary.
- *RandomizedTokenForwarding:* In each round, node u sends a piece of information to each of its neighbors which they don't have yet.

Conclusion

- Controlling harmful diffusions.
  - Give a $2d$ (or $O(\log n)$) approximation algorithm for centralized intervention strategies.
  - Show the existence (or non-existence) for decentralized intervention strategies, and give performance bound on the decentralized solutions with respect to optimal centralized solutions.
  - With the existence of risk behaviors, observe interesting phenomena and propose to give rigorous proofs.
- Enabling positive diffusions in dynamic networks.
  - Resource discovery: give almost tight bounds on converging time for both triangulation and two-hop walk processes.
  - Information dissemination in adversarial network: propose to devise efficient algorithms.