

# CASCADE PROCESSES IN SOCIAL NETWORKS

## a conjecture of Kempe, Kleinberg, Tardos

Sebastien Roch  
Microsoft Research

*based on: Mossel, Roch,*  
“On the Submodularity of Influence  
in Social Networks,” IEEE FOCS 2007



# outline of the talk

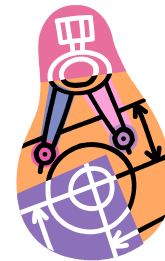
## background & motivation

- models of collective behavior
- influence maximization problem
- main result



## proof sketch

- definitions
- proof ideas
- antisense coupling



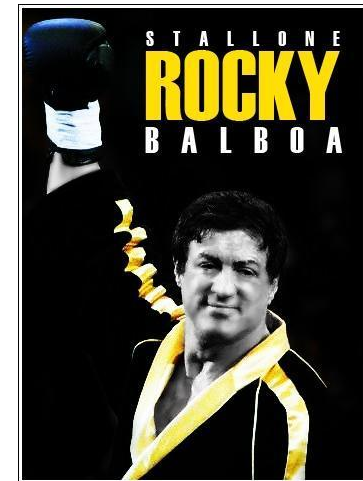
# PART I

## background & main result



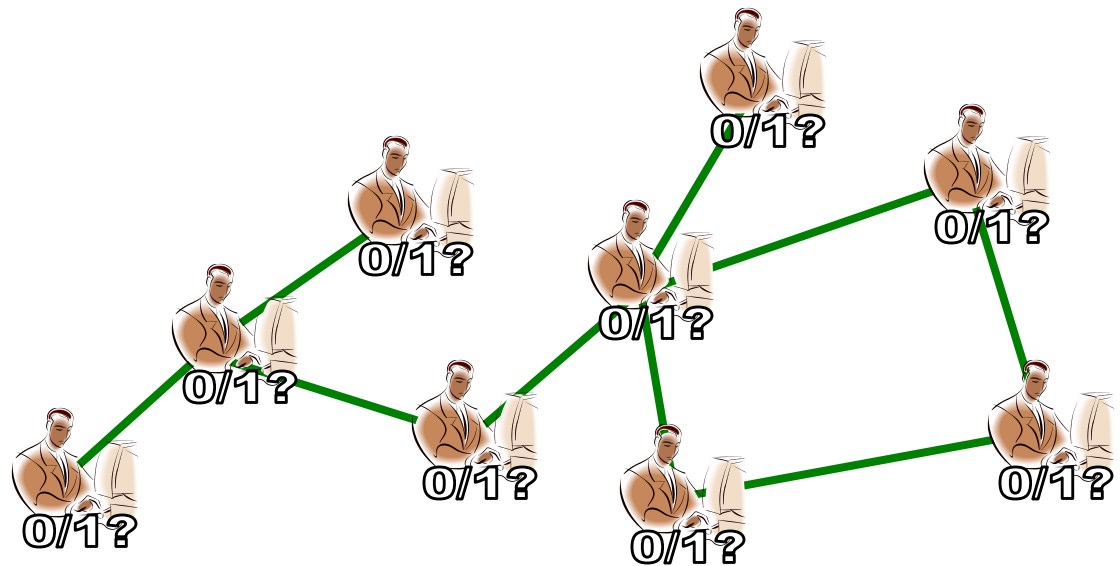
# models of collective behavior

- **examples:**
  - joining a riot
  - adopting a product
  - going to a movie
- **model features:**
  - binary decision
  - network structure
  - cascade effect

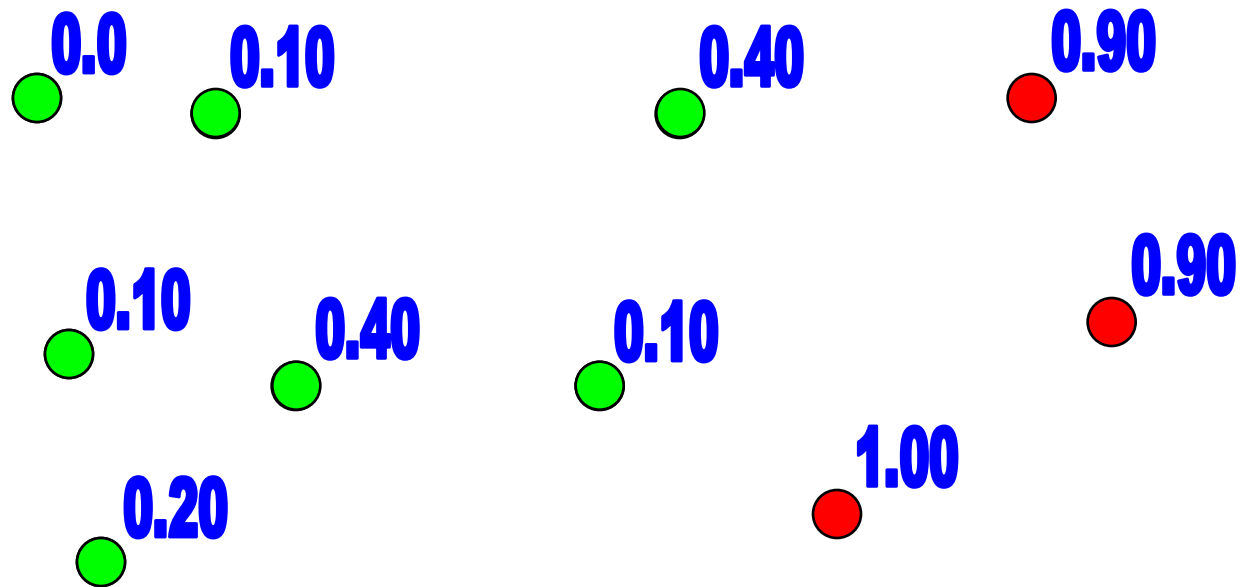


# viral marketing

- “traditional” marketing: target demographic
- referrals, word-of-mouth can be very effective (e.g. Hotmail, Gmail)
- **viral marketing** (e.g. [Domingos-Richardson KDD'01, SIGKDD'02])
  - goal: **mining the network value** of potential customers
  - how: **target a small set of trendsetters, seeds**



# Granovetter's threshold model



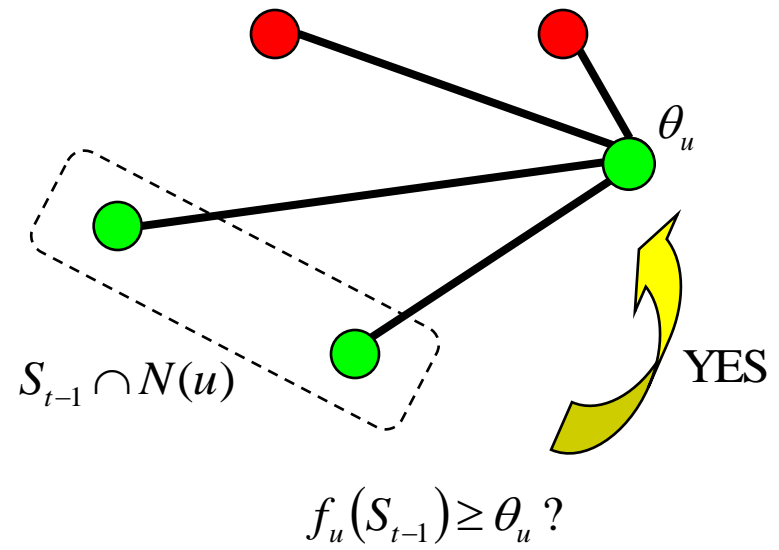
# generalized threshold model

- graph  $G=(V,E)$  with  $n$  nodes (“customers”)
- “infection” process  $(S_t)$  starts at given set  $S_0$
- **definition** [Kempe-Kleinberg-Tardos SIGKDD’03, ICALP’05] the **generalized threshold model** is defined as follows:
  - activation functions:
    - $\{f_u(\cdot)\}_{u \in V}$  where  $f_u$  depends only on neighbours  $N(u)$  of  $u$
  - threshold values:
    - $\{\theta_u\}_{u \in V}$  i.i.d. uniform in  $[0,1]$
  - dynamics:
    - at time  $t$ , set  $S_t$  to  $S_{t-1}$
    - and add all nodes  $u$  with

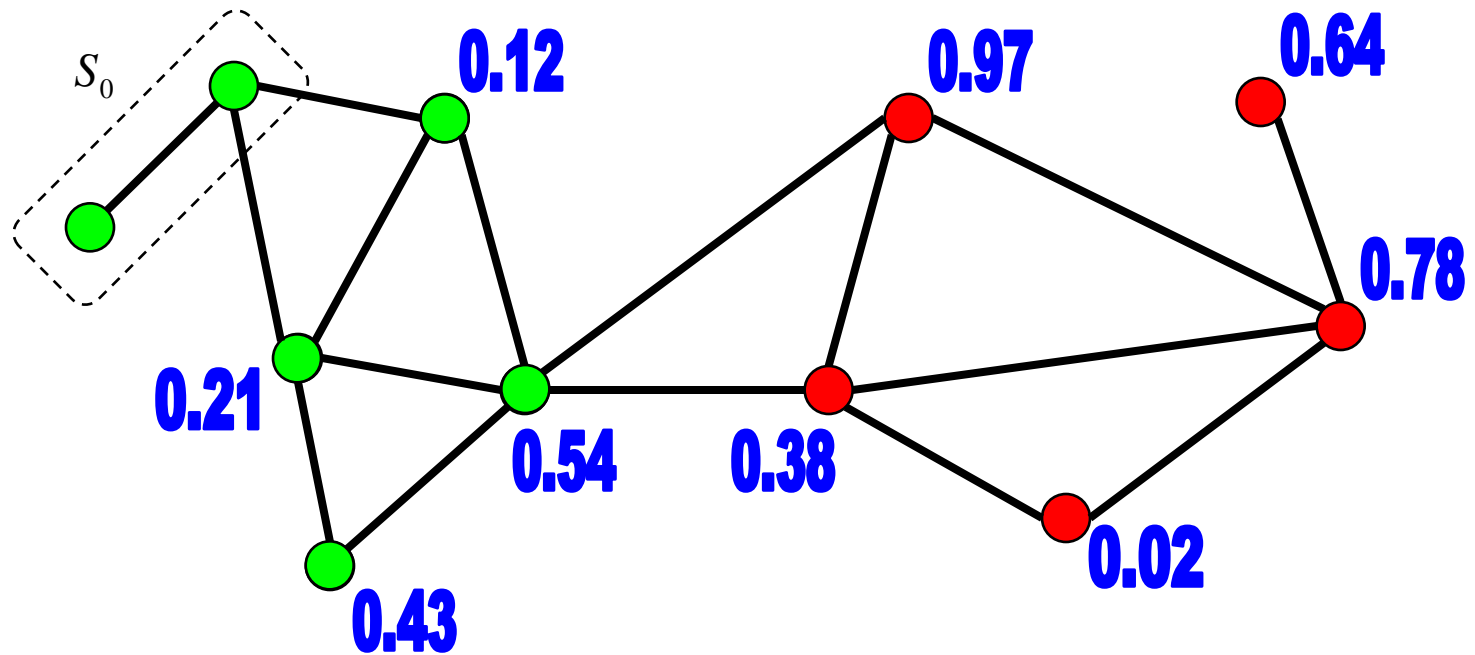
$$f_u(S_{t-1}) \geq \theta_u$$

● INFECTED

● NOT INFECTED



# diffusion on a network



stops after at most  $n-1$  steps

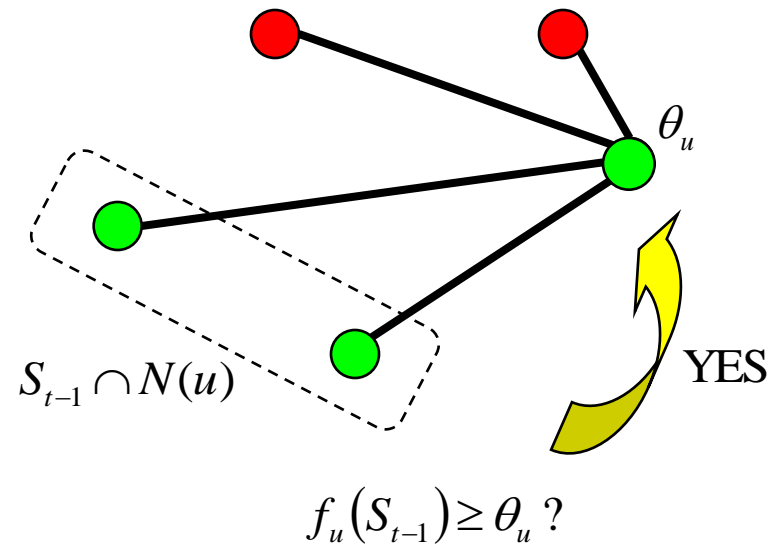
# generalized threshold model

- graph  $G=(V,E)$  with  $n$  nodes (“customers”)
- “infection” process  $(S_t)$  starts at given set  $S_0$
- definition** [Kempe-Kleinberg-Tardos SIGKDD’03, ICALP’05] the **generalized threshold model** is defined as follows:
  - activation functions:
    - $\{f_u(\cdot)\}_{u \in V}$  where  $f_u$  depends only on neighbours  $N(u)$  of  $u$
  - threshold values:
    - $\{\theta_u\}_{u \in V}$  i.i.d. uniform in  $[0,1]$
  - dynamics:
    - at time  $t$ , set  $S_t$  to  $S_{t-1}$
    - and add all nodes  $u$  with

$$F \circ f_u(S_{t-1}) \geq \theta_u^{-1}(U_u)$$

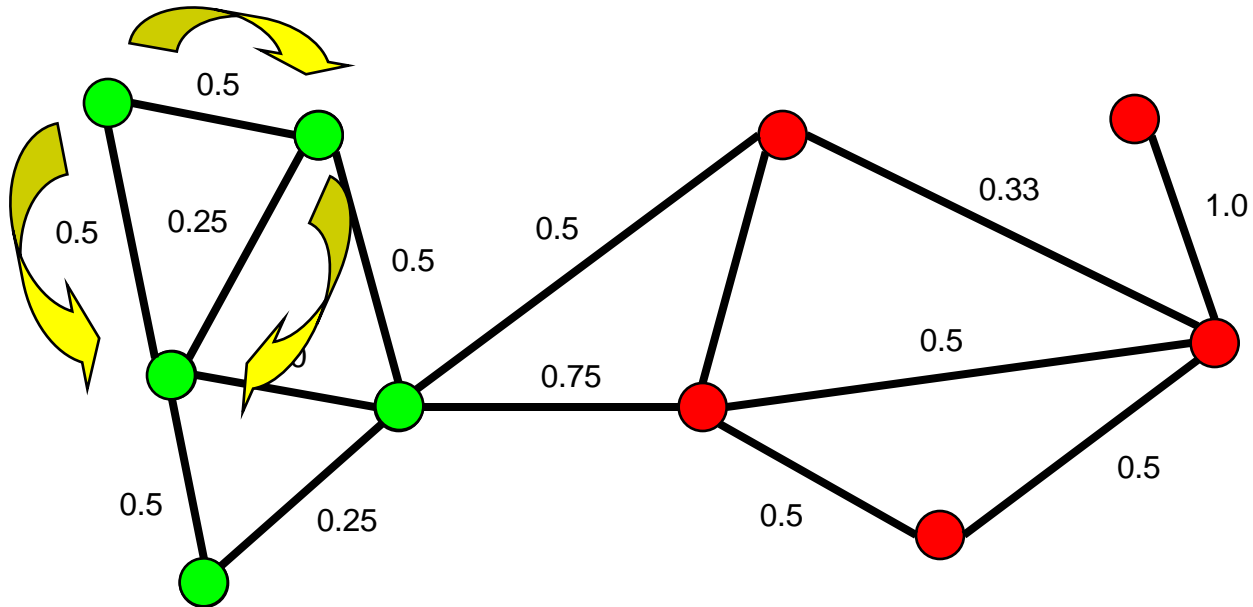
 INFECTED

 NOT INFECTED



# independent cascade model

- when a node is activated
  - it gets **one chance to activate** each neighbour
  - **probability of success** from  $u$  to  $v$  is  $p_{u,v}$



# influence maximization

- **definition** - the **influence**  $\sigma(S)$  given the initial seed  $S$  is the expected size of the infected set at termination

$$\sigma(S) = E_S [|S_{n-1}|]$$

- **definition** - in the **influence maximization problem (IMP)**, we want to find the seed  $S$  of fixed size  $k=k(n)$  that maximizes the influence

$$S^* = \arg \max \{ \sigma(S) : S \subseteq V, |S| = k \}$$

- standard heuristics: “degree centrality”, “distance centrality”
  - static, redundant, no guarantee
- **theorem** [KKT’03] - the IMP is **NP-hard**

# submodularity

- **definition** - a set function  $f : 2^V \rightarrow \mathbb{R}$  is **submodular** if for all  $A, B$  in  $V$

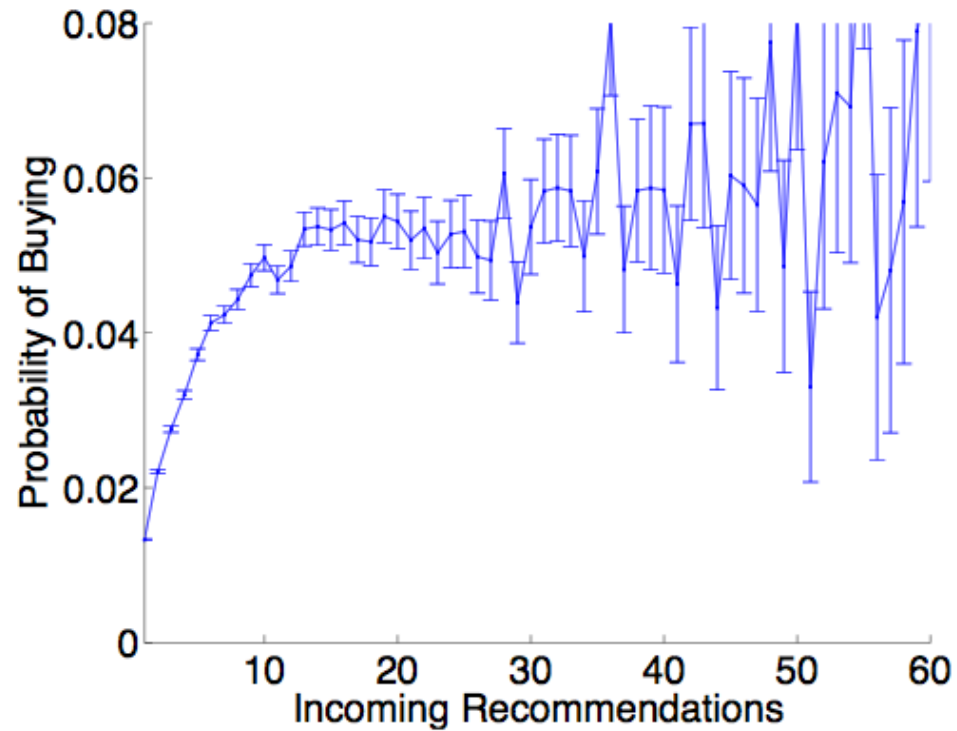
$$f(A) + f(B) \geq f(A \cap B) + f(A \cup B)$$

- **interpretation** - “discrete concavity” or “diminishing returns”, indeed submodularity equivalent to

$$\forall S \subseteq T, \forall v \in V, \quad f(T \cup \{v\}) - f(T) \leq f(S \cup \{v\}) - f(S)$$

- **example** -  $f(S) = g(|S|)$  where  $g$  is concave
- back to threshold models:
  - it is natural to assume that the **activation functions have diminishing returns**
  - supported by observations of [Leskovec-Adamic-Huberman'06] in the context of viral marketing

# Leskovec-Adamic-Huberman



# main result

- **theorem** [Mossel-R'07; first conjectured in KKT'03] - in the generalized threshold model, if all activation functions are monotone & submodular, then the **influence is monotone & submodular**, i.e.,

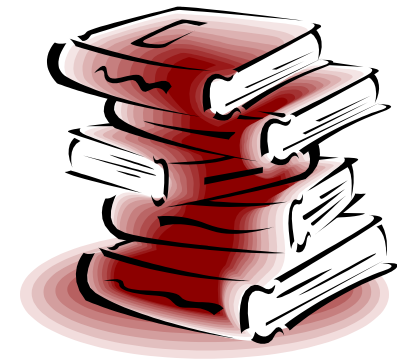
$$\sigma(S) = \mathbb{E}_S [|S_{n-1}|] \text{ is submodular''}$$

- **algorithmic corollary** [Mossel-R'07] - IMP admits a  $(1 - e^{-1} - \varepsilon)$ -**approximation algorithm** (for all  $\varepsilon > 0$ )
  - this follows from a general result on the approximation of submodular functions [Nemhauser-Wolsey-Fisher'78]
- known special cases [KKT'03,'05]:
  - linear threshold model, independent cascade model
  - “normalized” submodular threshold model

$$\forall S \subseteq T, \frac{f_u(S \cup \{v\}) - f_u(S)}{1 - f_u(S)} \geq \frac{f_u(T \cup \{v\}) - f_u(T)}{1 - f_u(T)}$$

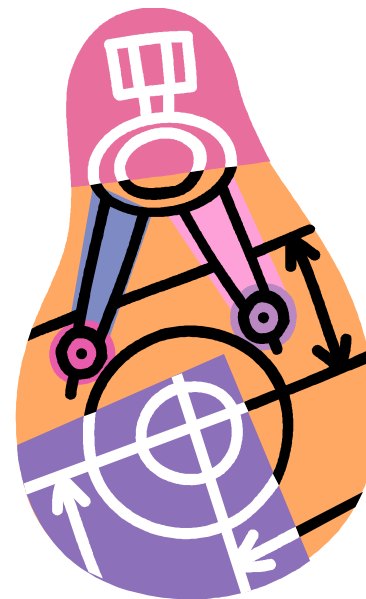
# related work

- **sociology, economics**
  - threshold models: [Granovetter'78], [Morris'00]
  - cascades: [Watts'02]
  - games, diffusion of innovations: [Ellison'93], [Young'02]
- **CS, data mining**
  - viral marketing: [KKT'03,'05], [Domingos-Richardson'01,'02]
  - recommendation networks, blogs: [Leskovec-Singh-Kleinberg'05], [Leskovec-Adamic-Huberman'06]
- **probability theory**
  - percolation
  - voter model, contact process
  - Markov random fields



# PART II

## proof sketch



# coupling: monotonicity

- fix activation functions (mono.); sets  $A$  included in  $B$ , **want to prove:**

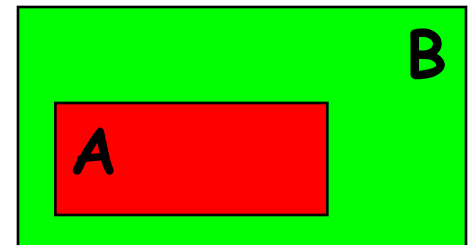
$$\sigma(B) \geq \sigma(A)$$

- consider 2 processes:
  - $(A_t)$  started at  $A$  with  $\theta$ -vector =  $\theta^A$
  - $(B_t)$  started at  $B$  with  $\theta$ -vector =  $\theta^B$
- proof idea** - take  $\theta^A = \theta^B$ , then

$$A_t \subseteq B_t \quad \forall t$$

and

$$|A_{n-1}| \leq |B_{n-1}|$$



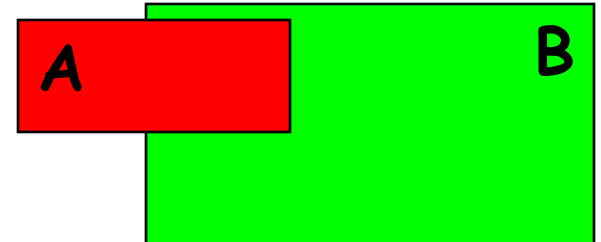
# coupling: submodularity

- fix activation functions (subm.); arbitrary sets  $A, B$ , **want to prove:**

$$\sigma(A) + \sigma(B) \geq \sigma(A \cap B) + \sigma(A \cup B)$$

- consider 4 processes:

- $(A_t)$  started at  $A$
- $(B_t)$  started at  $B$
- $(I_t)$  started at  $A \cap B$  ["intersection" process]
- $(U_t)$  started at  $A \cup B$  ["union" process]



- proof idea** - it suffices to **couple the 4 processes** such that

$$I_t \subseteq A_t \cap B_t \quad (1) \quad U_t \subseteq A_t \cup B_t \quad (2) \quad \forall t$$

indeed, then

$$|A_{n-1}| + |B_{n-1}| = |A_{n-1} \cap B_{n-1}| + |A_{n-1} \cup B_{n-1}| \geq |I_{n-1}| + |U_{n-1}|$$

# high-level proof idea

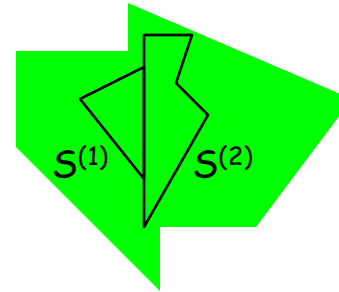
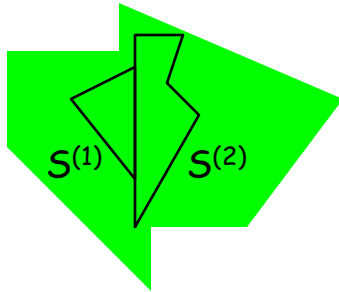
- our goal:

$$I_t \subseteq A_t \cap B_t \quad (1) \quad U_t \subseteq A_t \cup B_t \quad (2)$$

- **antisense coupling**
  - obvious way to couple: use same  $\theta_u$ 's for all 4 processes
  - satisfies (1) but not (2)
  - “antisense”: using  $\theta_u$  for ( $A_t$ ) and  $(1-\theta_u)$  for ( $B_t$ ) “**maximizes union**”
  - we combine both couplings

# piecemeal growth

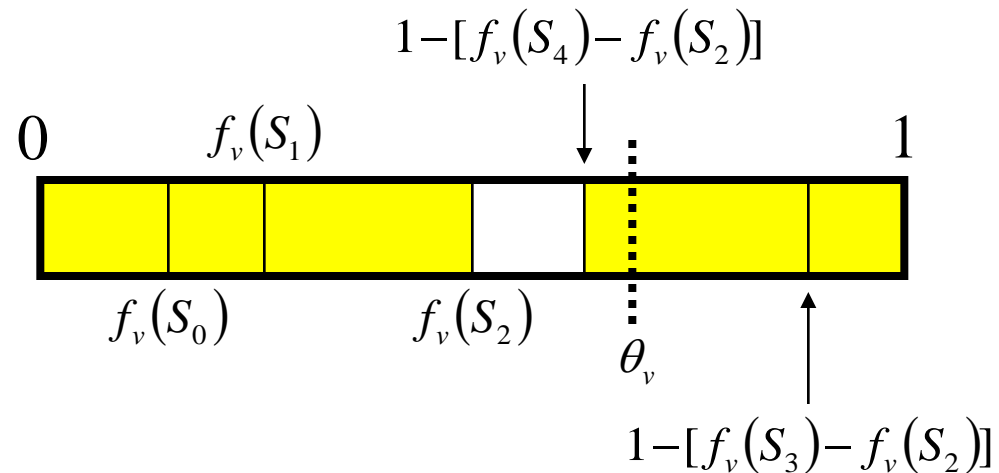
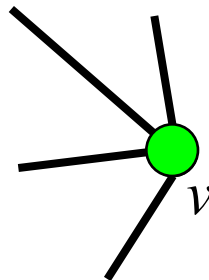
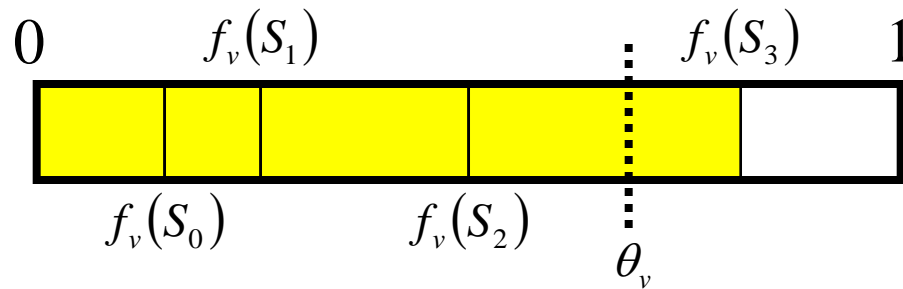
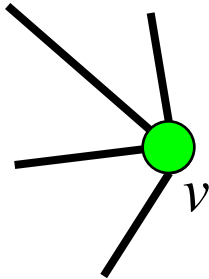
- **partition** of  $S$ :  $S^{(1)}, S^{(2)}$



- **lemma** - the final sets are the same

# antisense coupling

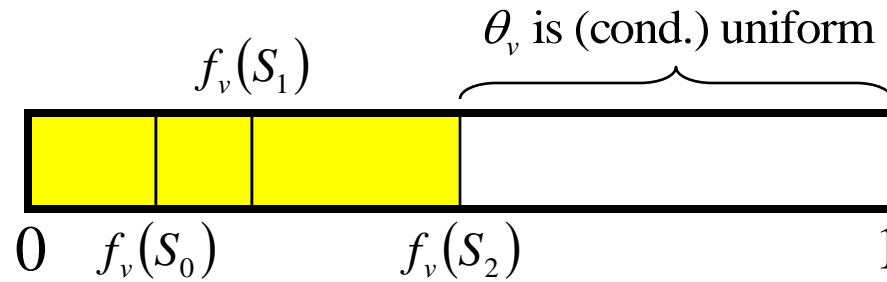
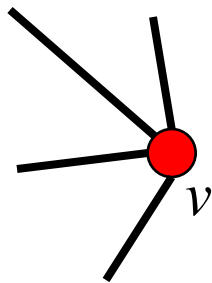
- disjoint sets:  $S^{(1)}, S^{(2)}$



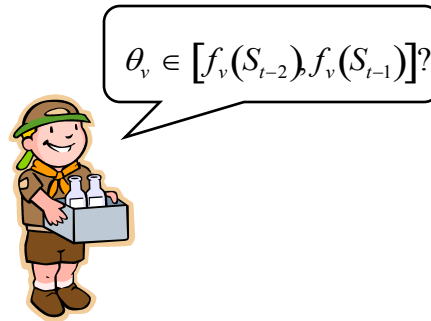
- lemma** - the final sets have the same distribution

# need-to-know

- proof of lemma



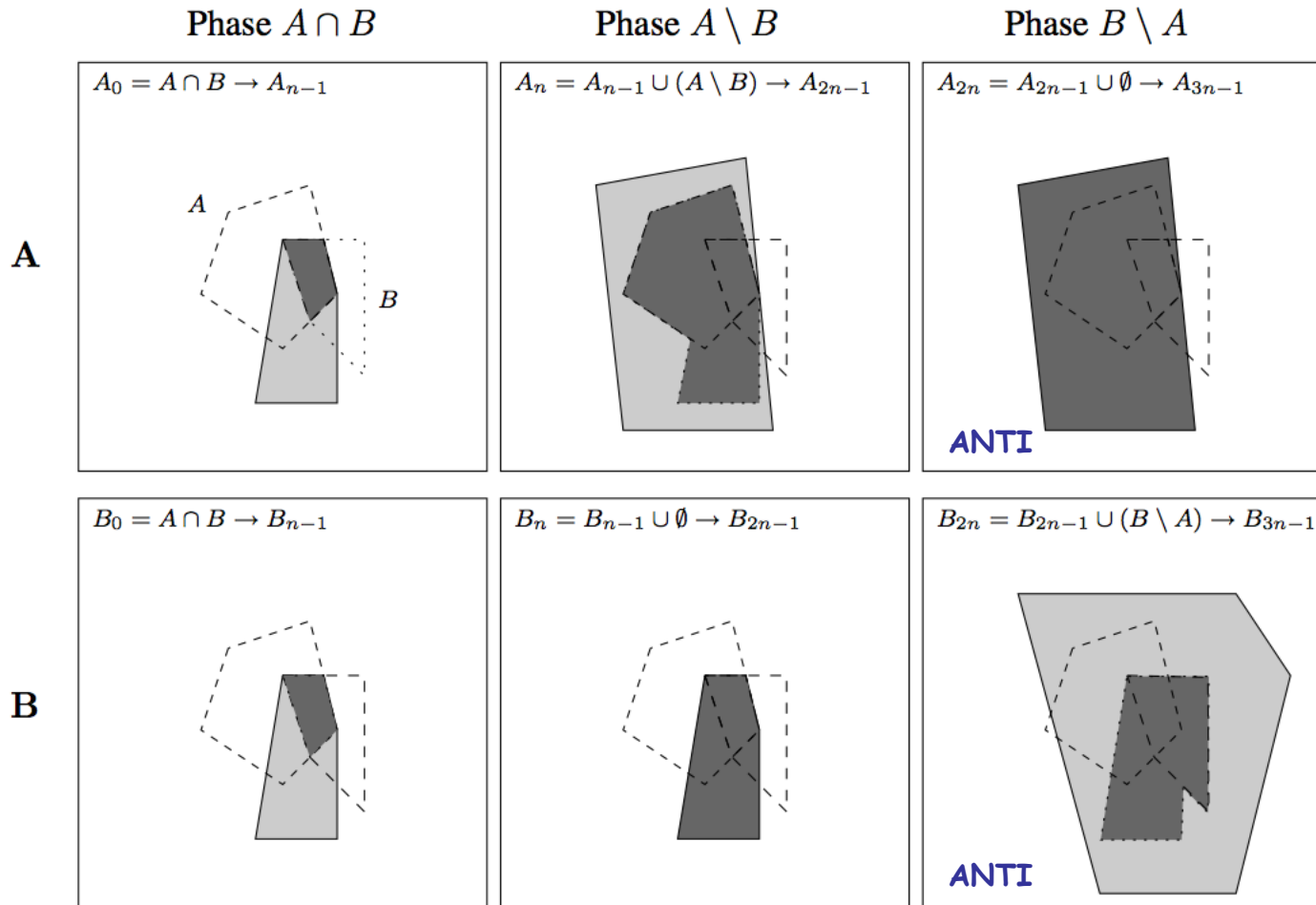
simulation 1



simulation 2



# proof I



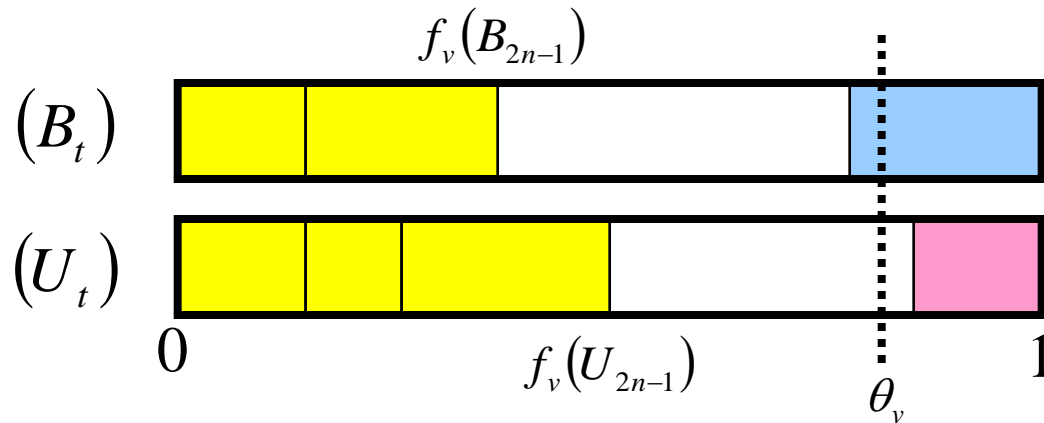
# proof II

- our goal:

$$I_t \subseteq A_t \cap B_t \quad (1) \quad U_t \subseteq A_t \cup B_t \quad (2)$$

- (1) is trivial
- up to time  $2n-1$ , we have  $A_t = U_t$  so (2) holds in first two phases
- at time  $2n$ , by **monotonicity** and **submodularity**

$$f_v(B_{2n}) - f_v(B_{2n-1}) \geq f_v(U_{2n}) - f_v(U_{2n-1})$$



conclude  
by induction

# general result

- we have proved:

**theorem** [Mossel-R'07] - in the generalized threshold model, if all activation functions are submodular, then for any monotone, submodular function  $w$ , the **generalized influence**

$$\sigma_w(S) = E_S[w(S_{n-1})]$$

is submodular

# final remarks

- inference
- non-submodular case
- outbreak detection [Leskovec et al.'07]
- monetizing social networks [Hartline-Mirrokn-Sundarajan'07]



**thank**  
**you**