# On the Submodularity of Influence in Social Networks

Elchanan Mossel & Sebastien Roch

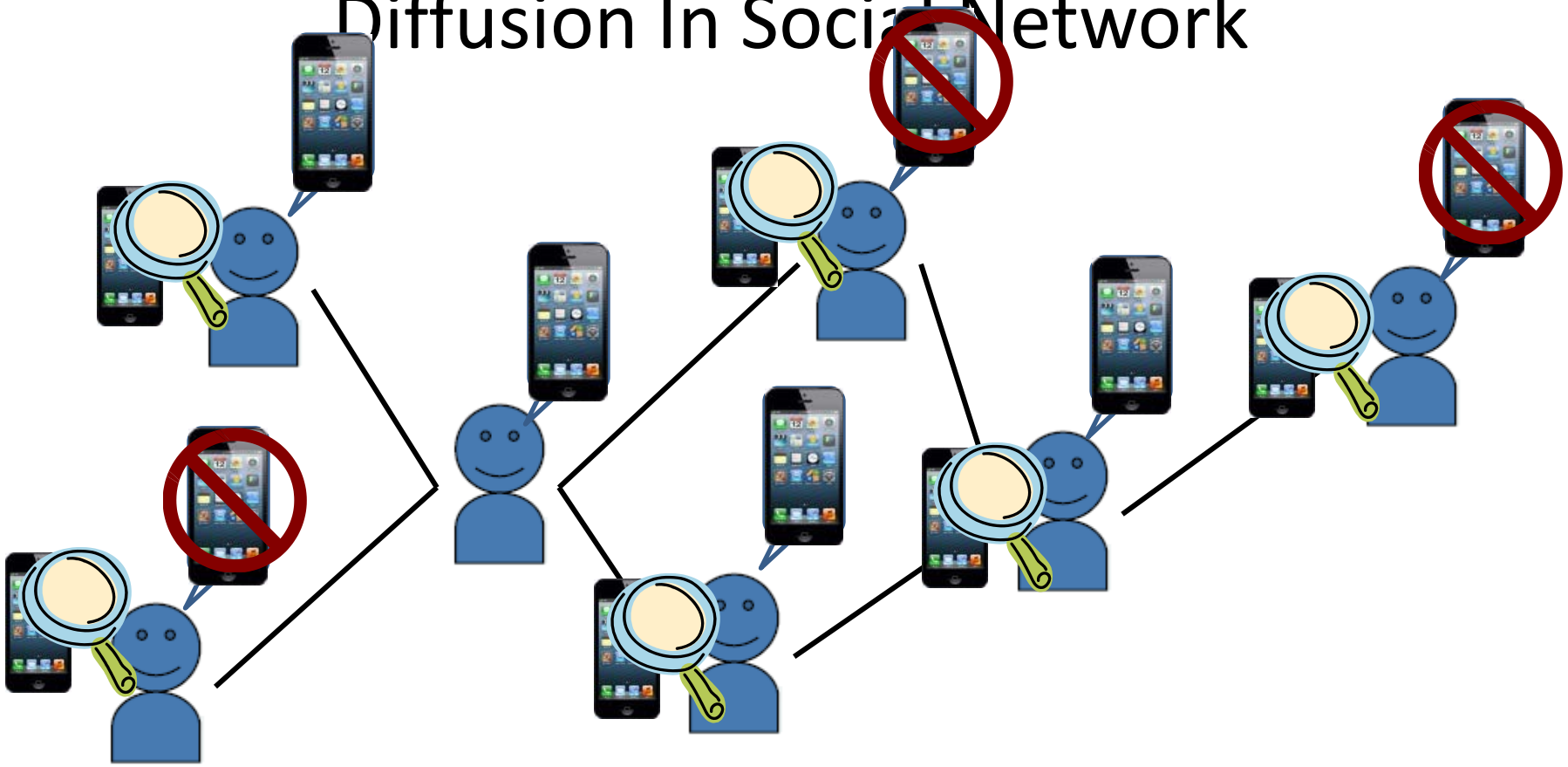
STOC07

Speaker: Xinran He
Xinranhe1990@gmail.com

# Social Network

- Social network as a graph
  - –Nodes represent individuals.
  - –Edges are social relations with different strengths:
    - Neighbors, Coworkers relation in real life
    - Virtual Friendship in Facebook
    - Follower-Followee relations in Twitter

# Diffusion In Social Network



- The adoption of new products  can propagate in the social network → **Diffusion in the social network**
- Information, rumors, innovation, ……

# Influence Maximization

- Influence maximization:  Find k people that generates the largest influence spread (i.e. expected number of activated nodes) [KKT 2003]

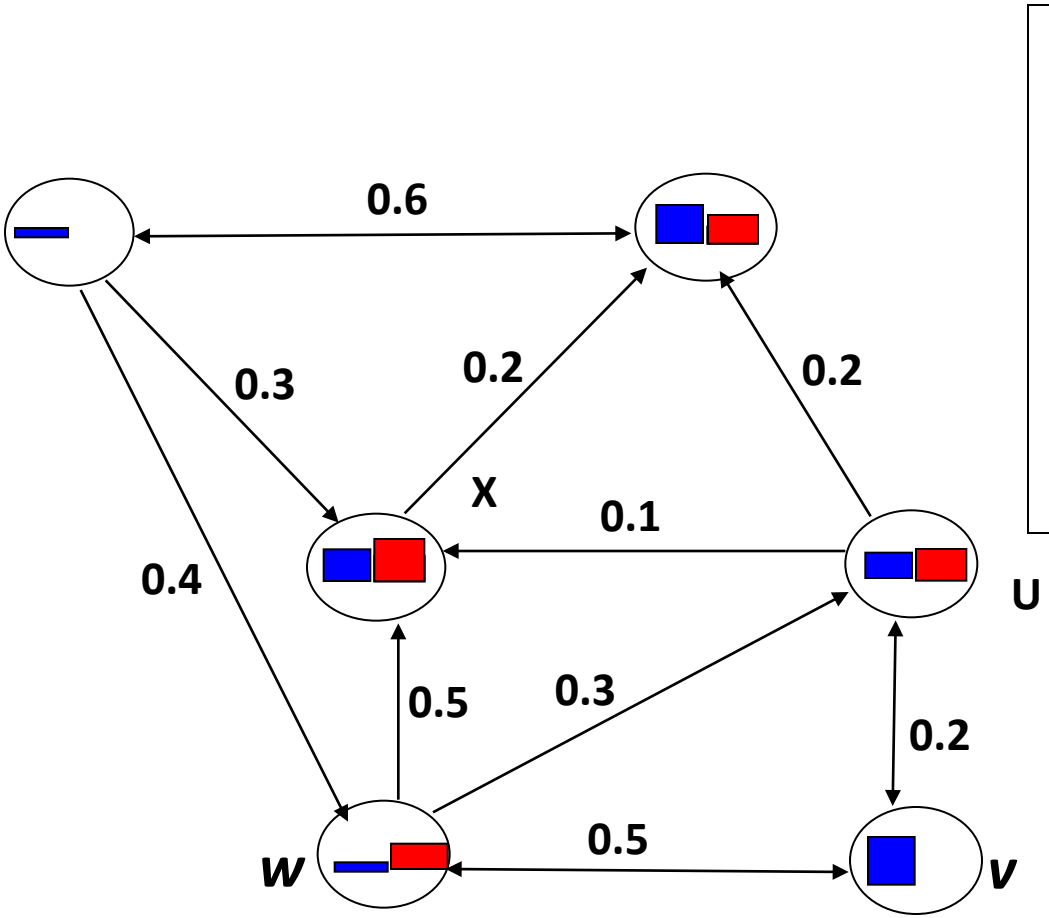
# Linear Threshold Model

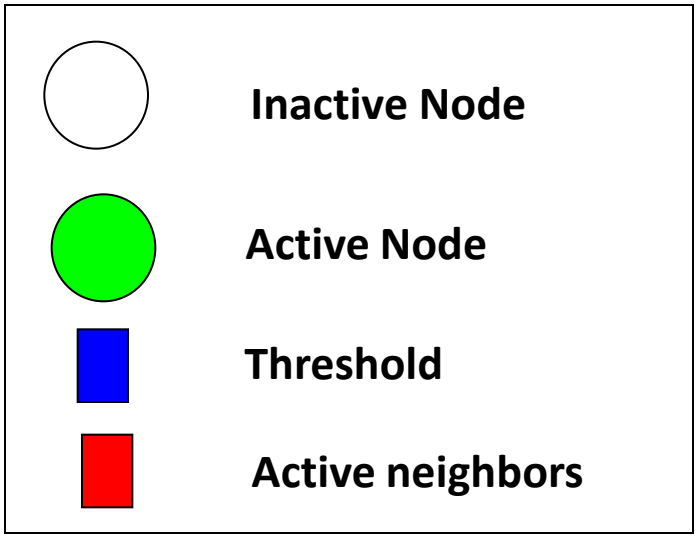
- Given a social network with edge weight $w_{uv}$ and a set of Initially active individuals S as seed.
- Every individual independently chooses a threshold $\Theta_v$ uniformly in [0,1].
- At any step t later, still inactive nodes become activated if

$$\sum_{u \in N_v} w_{uv} \geq \theta_v$$

  where $N_v$ is the set of activated direct neighbors of v.
- The diffusion ends when no more nodes are activated.
- The influence spread $\sigma(S)=E[|P_{end}||S]$, is the expected number of active nodes when the diffusion process ends.

# Linear Threshold Example




***Stop!***

**Step 4**

# Influence Maximization

- Find a seed set S, $|S| \leq k$, $\sigma(S)$ is maximized.
- Influence Maximization Problem is NP-hard under linear threshold model[Kempe et.al 2003].
- We have to solve it approximately.
- Main tool for analysis

**Theorem**: The greedy algorithm is a 1-1/e approximation for maximizing monotone and submodular set functions[Nemhauser/Wolsey 1978].

# Submodular & Monotone

- A set function f: $2^V \rightarrow$ R is **monotone** if

$$f(S) \leq f(T), \text{ for all } S \subseteq T \subseteq V$$

- A set function f: $2^V \rightarrow$ R is **submodular** if

$$f(S) + f(T) \geq f(S \cap T) + f(S \cup T)$$

$$\text{for all } S, T \subseteq V$$

# Submodularity

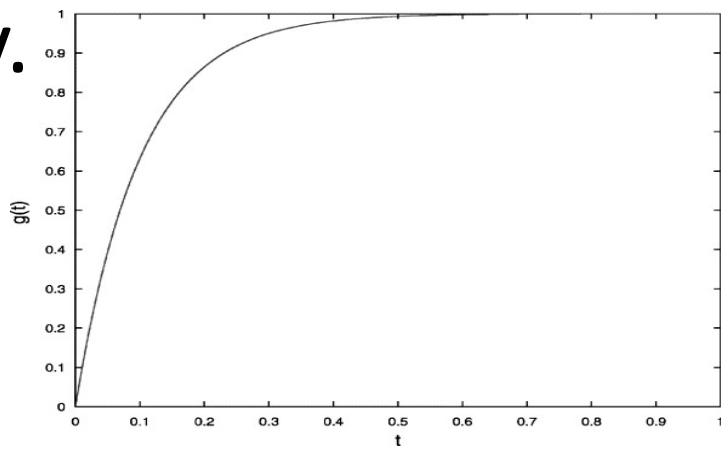
- A function set f is submodular if

$$f(S)+f(T)\geq f(S\cap T)+f(S\cup T), \text{for all } S,T\subseteq V$$

- Or equivalently

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

- Submodularity can be considered as diminishing return property.



# Submodularity: Examples

- Maximum coverage problem:

  Given a collection of sets $\mathbb{S}$={$S_1$,…,$S_m$} and a number k, find $S' \subseteq S, |S'| \leq k$, maximize σ(S')=$\left| \bigcup_{S_i \in S'} S_i \right|$.
  σ is **submodular**.

- The influence spread σ under the linear threshold model is **submodular**[Kempe et.al 2003]. → Influence Maximization Problem under linear Threshold model can be solved approximately.

# General Threshold Model

Linear Threshold Model: $\sum w_{uv} \geq \theta_v$

**General Threshold Model: $f_v(S) \geq \theta_v$**

$f_v(S)$ : activation function of node v over S. S is the set of already activated nodes.

- General Threshold model is generalization of many diffusion models:

$$f_v(S) = \begin{cases} \sum_{u \in N_v} w_{uv} & \text{Linear Threshold Model [KKT 2003]} \\ 1 - \prod_{u \in N_v} (1 - p_{uv}) & \text{Independent Cascade Model [KKT 2003]} \\ 1 - \prod_{i=1}^{r} (1 - p_v(\omega_i, S_{i-1})) & \text{Decreasing Cascade Model [KKT 2005]} \\ \dots & \dots \end{cases}$$

# General Threshold Model(2)

For Linear Threshold model, the influence spread σ(S) is **submodular** [KKT 2003].

**Conjecture**: Under the general threshold model with monotone and submodular $g_v(\cdot)$, $\sigma(S)$ is monotone and submodular [KKT 2003].

Yes!

Is it true?

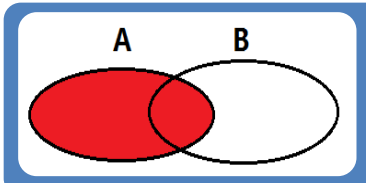
# Main Result

**Theorem:** Under the general threshold model with monotone and submodular $f_v$, $\sigma(S)$ is monotone and submodular [Mossel/Roch 2007].

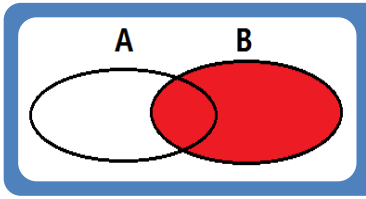
**Corollary:** The greedy algorithm is a (1-1/e) approximation to solve the influence maximization problem under general threshold model.

# Proof: General Idea(1)

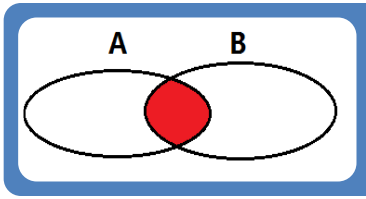
- By coupling four diffusion process:



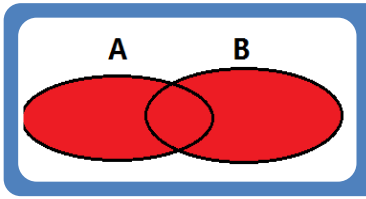

$\mathbb{A}=\{A_0=S, A_1, A_2, \ldots, A_{end}\}$




$\mathbb{B}=\{B_0=T, B_1, B_2, \ldots, B_{end}\}$




$\mathbb{C}=\{C_0=S\cap T, C_1, C_2, \ldots, C_{end}\}$

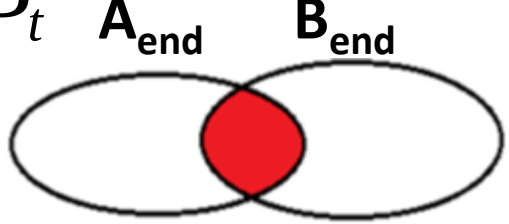



$\mathbb{D}=\{D_0=S\cup T, D_1, D_2, \ldots, D_{end}\}$

- Such that $C_t \subseteq A_t \cap B_t$ and $D_t \subseteq A_t \cup B_t$

# Proof: General Idea(2)

If $C_t \subseteq A_t \cap B_t$ and $D_t \subseteq A_t \cup B_t$




Then $|A_{end}| + |B_{end}|$

$$\geq |A_{end} \cap B_{end}| + |A_{end} \cup B_{end}|$$

$$\geq |C_{end}| + |D_{end}|$$

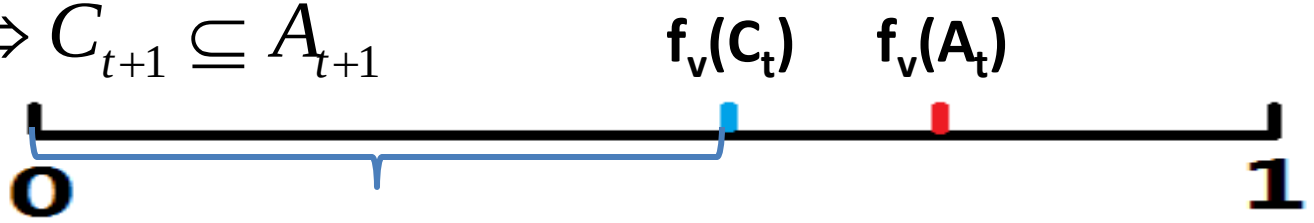
Then taking expectation, we have

$$\sigma(S) + \sigma(T) \geq \sigma(S \cap T) + \sigma(S \cup T)$$

$$C_t \subseteq A_t \cap B_t$$

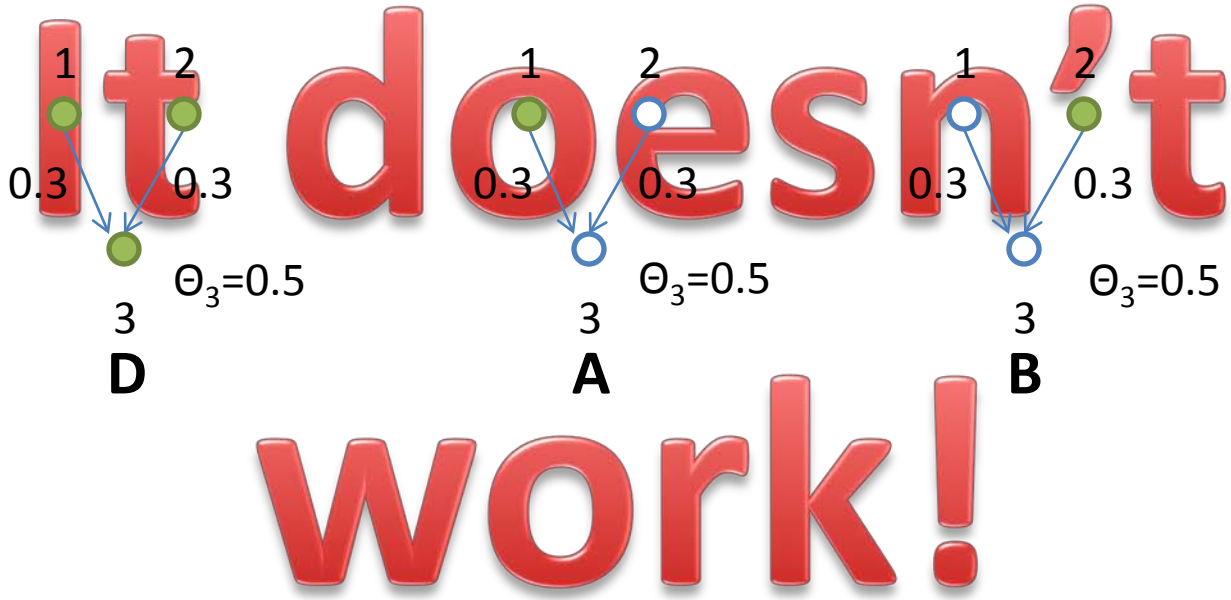
- Couple the four processes with the same thresholds $\theta_v$.
- Show $C_t \subseteq A_t, C_t \subseteq B_t$ by induction.
  - Base Case: $C_0 = S \cap T \subseteq S = A_0$
  - Assume $C_t \subseteq A_t$ .
  - For a node v still inactive at step t, we have $f_v(C_t) \leq f_v(A_t)$ . Therefore if v is activated in step t+1 in C, it must also be activated in A.

$$\Rightarrow C_{t+1} \subseteq A_{t+1}$$



# $D_t \subseteq A_t \cup B_t$ :First Attempt

- Let's try the same coupling method for $D_t \subseteq A_t \cup B_t$ .



# Antisense Coupling

- Then how could we keep $D_t \subseteq A_t \cup B_t$ ?

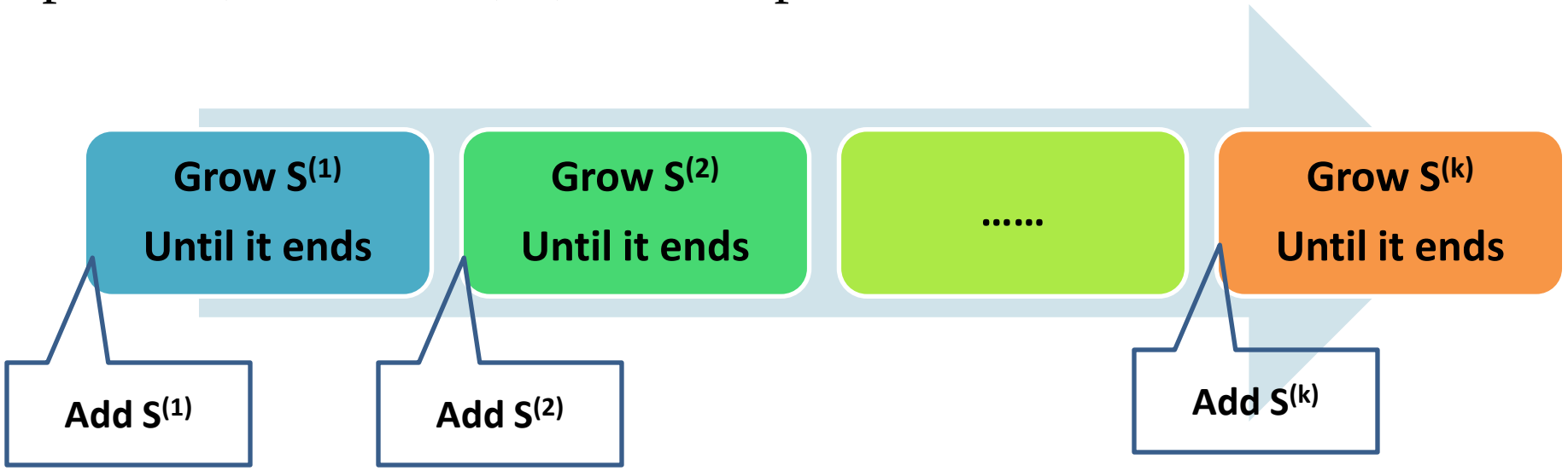
- Intuitively, using Θ for activation of S and 1-Θ for activation of T will maximize their union.

**Antisense Coupling**

**→Piecemeal Growth**

# Piecemeal Growth

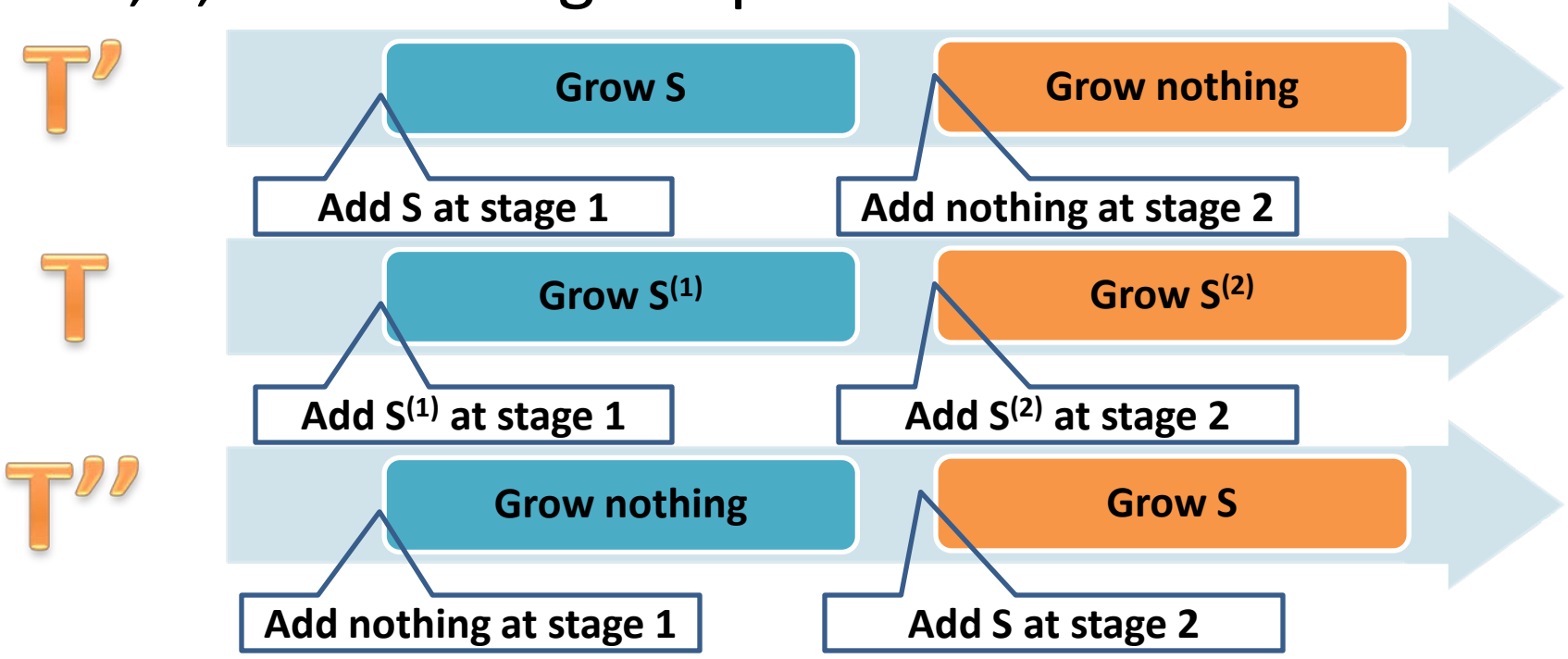
Define $P = P(S^{(1)},...,S^{(k)})$ as the the piecemeal growth diffusion process, where $S^{(1)},...,S^{(k)}$ is a partition of seed set $S$.




**Lemma:** The distribution over the activated node set at the end of original process with seed set S and the piecemeal growth process $P(S^{(1)},...,S^{(k)})$ is identical.

# Piecemeal Growth: Proof

- By coupling three piecemeal growth processes T', T, T'' and original process S with same θ.




$$T''_s \subseteq T_s \subseteq T'_s \text{ and } T'_{end} = T''_{end} = S_{end}$$

so that $S_{end} = T_{end}$

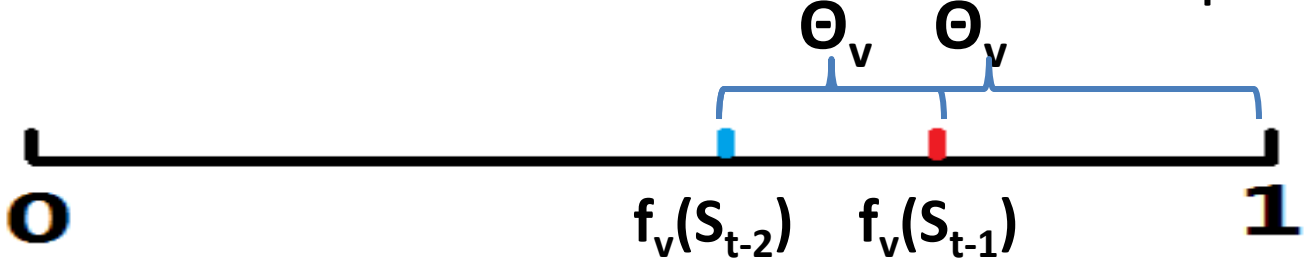
# Need-to-know Representation(1)

- Consider the diffusion in a different way:

  **Need-to-know Representation**.

- **Principle of Deferred Decisions:** We don't decide all thresholds at the beginning; instead we reveal the value of thresholds whenever needed.

- For example: if node v is inactive at step t-1, we only want to know whether it is activated at step t.



# Need-to-know Representation(2)

**Lemma:** The following process is equivalent to the original one:

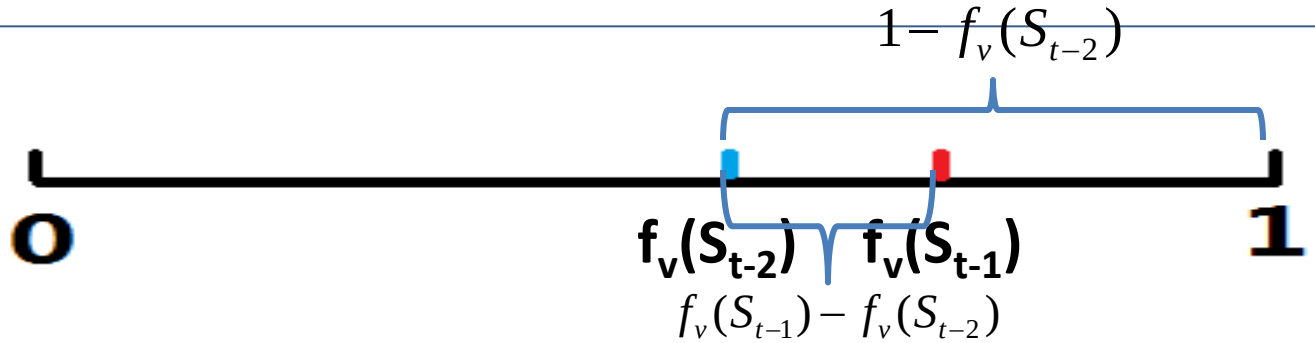
1. Initialize $S_0 = S$

2. At step $1 \le t \le n-1$, we initialize $S_t = S_{t-1}$ and for each still inactive node $v$

- With probability $\dfrac{f_v(S_{t-1}) - f_v(S_{t-2})}{1 - f_v(S_{t-2})}$, $v$ becomes activated

  and we pick $\theta_v$ uniformly in $[f_v(S_{t-2}), f_v(S_{t-1})]$.

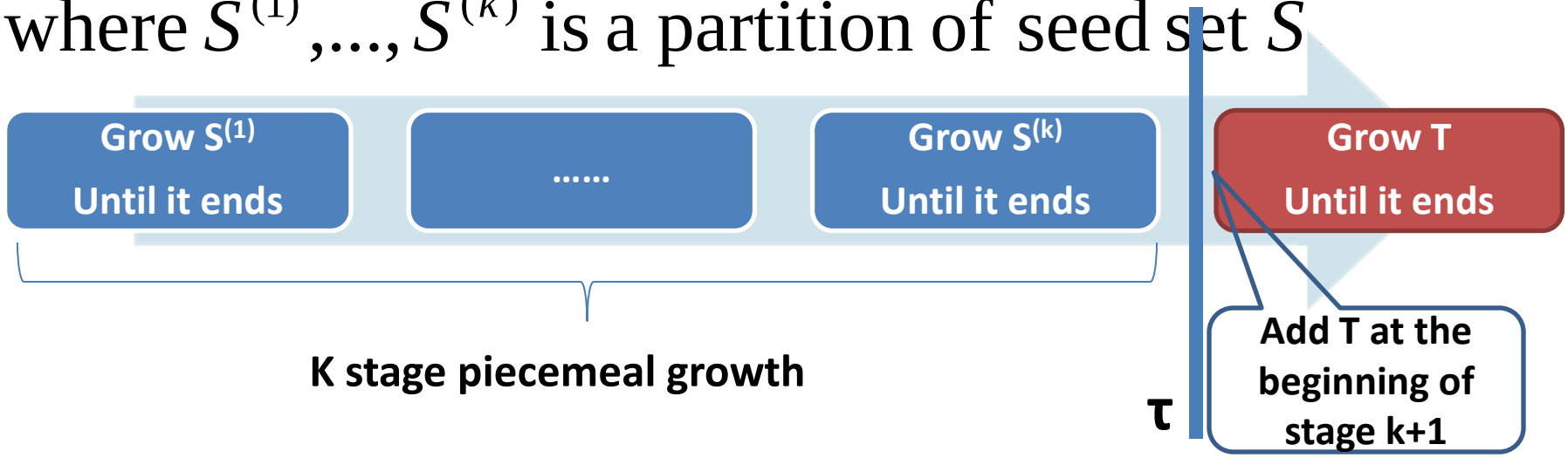
- Otherwise we do nothing.



# Antisense Coupling(1)

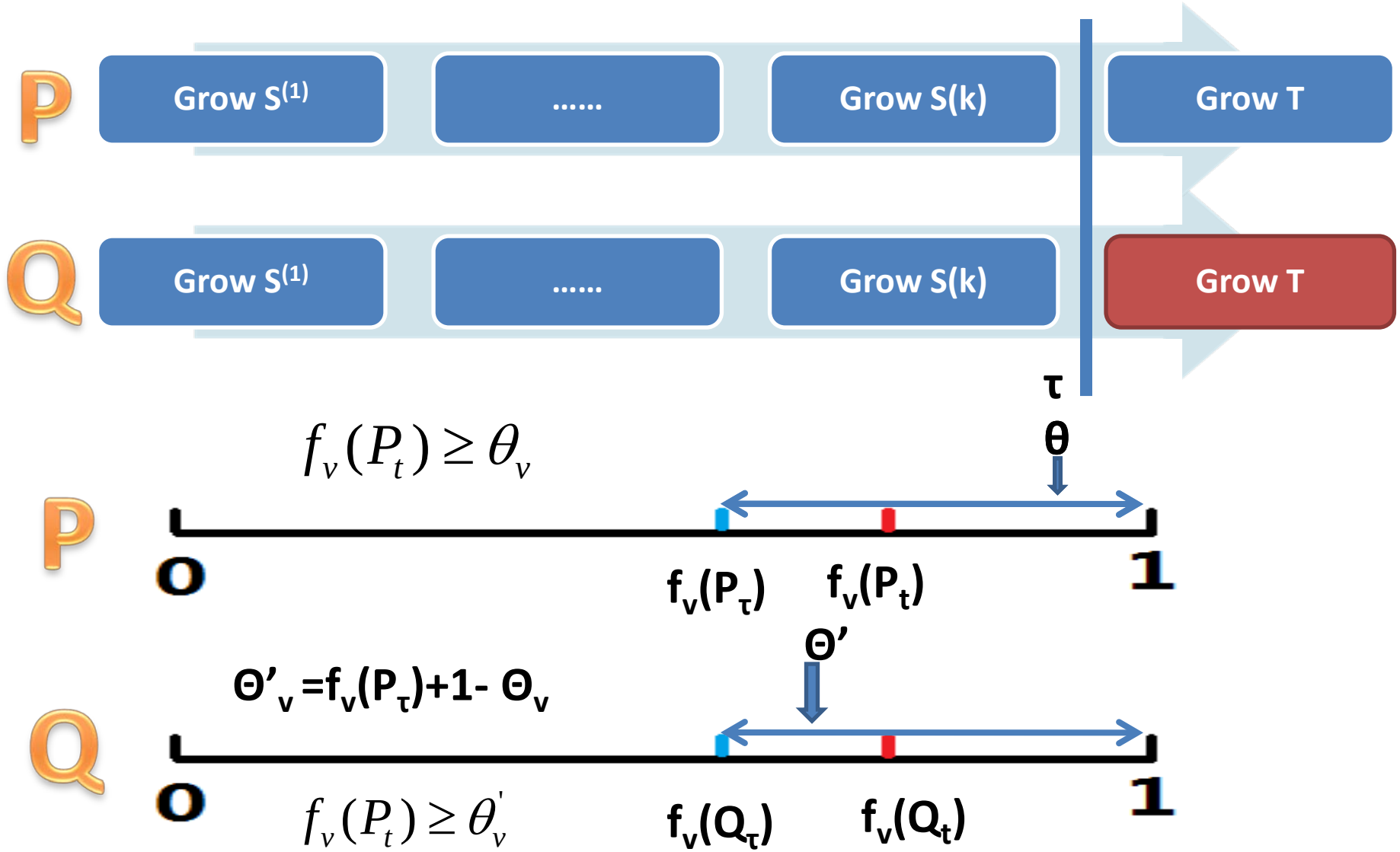
Define the antisense diffusion $P = P(S^{(1)},...,S^{(k)};T)$

where $S^{(1)},...,S^{(k)}$ is a partition of seed set $S$




Any step t in the final stage, activate nodes under the condition $f_v(P_t) \geq f_v(P_\tau) + 1 - \theta_v$.

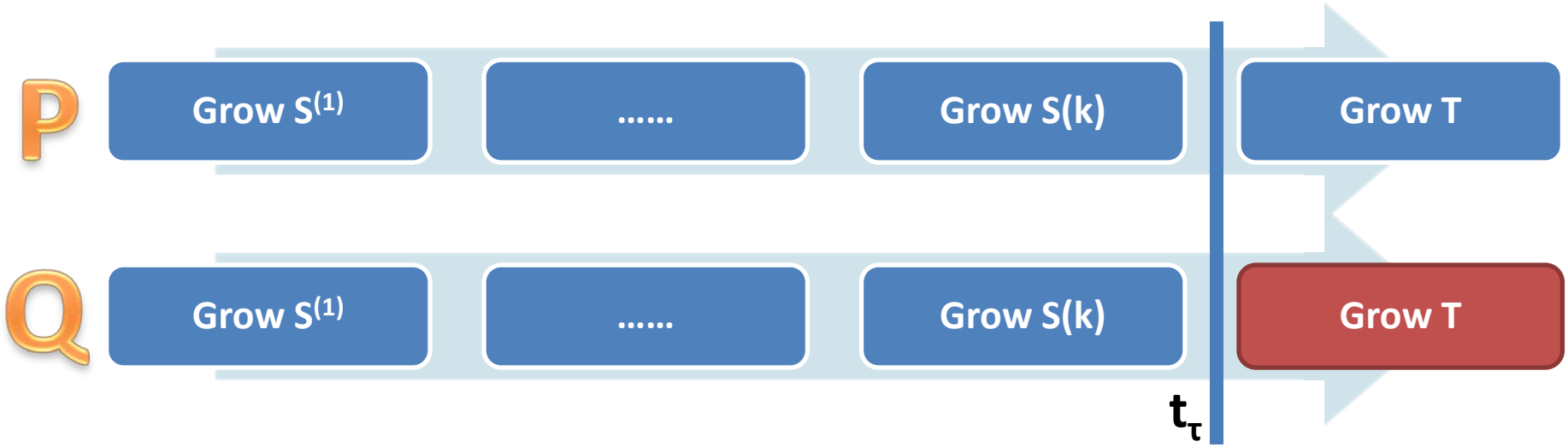
# Antisense Coupling(2)



# Antisense Coupling(3)


**Lemma:** The distributions over the activated node set at the end of the piecemeal growth process $P(S^{(1)},\ldots,S^{(k)};T)$ and the antisense diffusion process $Q(S^{(1)},\ldots,S^{(k)};T)$ are identical.

# Antisense Coupling: Proof(1)




- From **Need-to-know Representation** point of view:

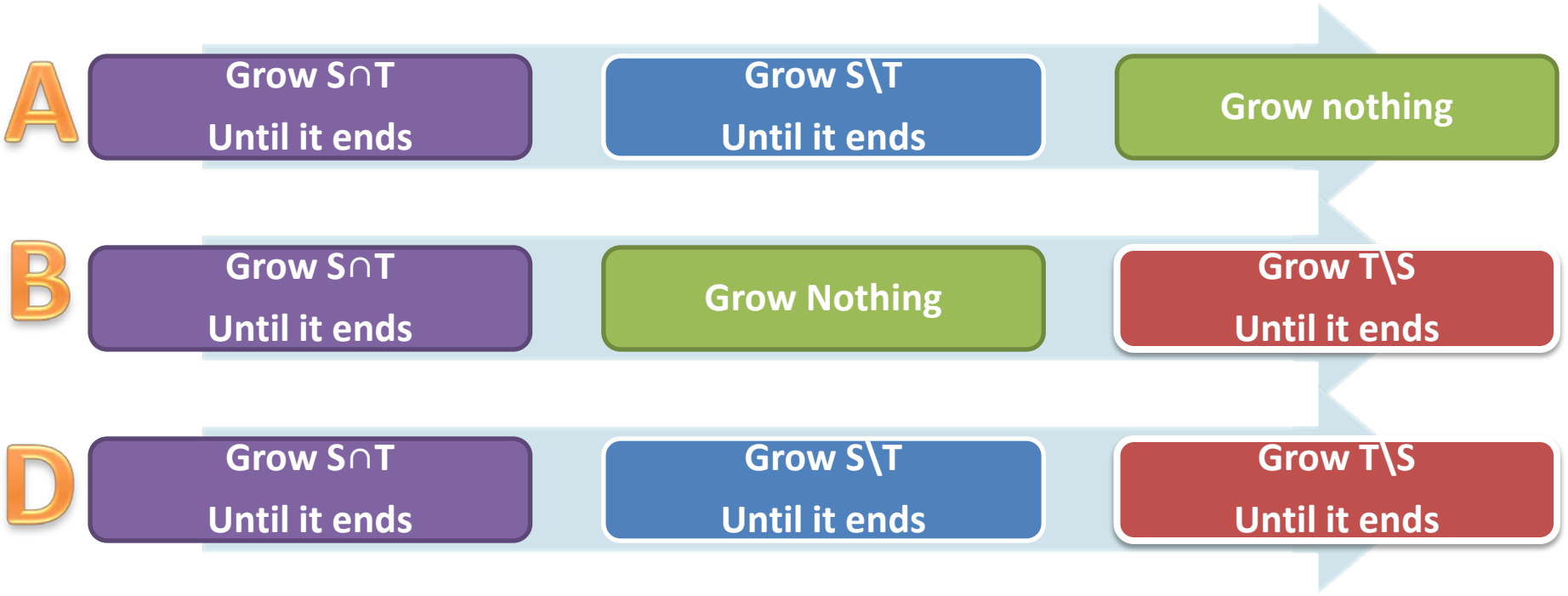
  For any node $v$ still inactive at time $t = \tau$, we have

  $\theta_v$ uniformly distributed in $[f_v(P_\tau), 1] = [f_v(Q_\tau), 1]$

# Antisense Coupling: Proof(2)

- Then for any still inactive node, we pick its $\Theta_v$ uniformly in $[f_v(P_\tau),1]$.
- We define $\Theta'_v = f_v(Q_\tau)+1-\Theta_v$ .
- Since $\Theta_v$ and $\Theta'_v$ have the same distribution, the final stage in growing T in P and Q is identical.
- Therefore $P_{end}$ and $Q_{end}$ have the same distribution.

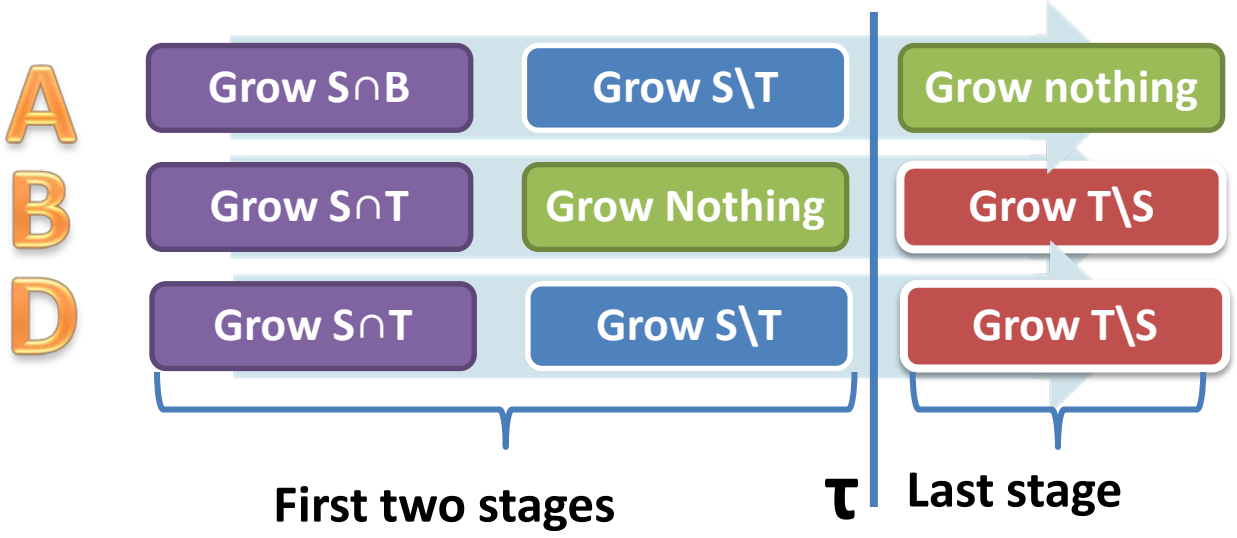
# Coupling: Overview




$D_t \subseteq A_t \cup B_t$ for any step t in all three stages

# Coupling: First two stages

- $A_t$=$D_t$ for all t in the first two stages.
- Therefore $D_t \subseteq A_t \cup B_t$ for all steps t in the first two stages.
- We will show $D_t \subseteq A_t \cup B_t$ for any step in final stage.



# Coupling: Antisense Coupling

- We first prove $D_t \setminus D_\tau \subseteq B_t \setminus B_\tau$ for any step in the final stage by induction on t.
- Base case:

$$D_{\tau+1} \setminus D_\tau \subseteq B_{\tau+1} \setminus B_\tau$$

- Because:

$$D_{\tau+1} = D_\tau \cup (T \setminus S)$$

$$B_{\tau+1} = B_\tau \cup (T \setminus S)$$

$$B_\tau \subseteq D_\tau$$

# Coupling: Antisense Coupling

- Assume $D_t \setminus D_\tau \subseteq B_t \setminus B_\tau$ .
- We need to show that $D_{t+1} \setminus D_\tau \subseteq B_{t+1} \setminus B_\tau$ .

**Lemma:** For any $S \subseteq S'$ and $T \subseteq T'$ and submodular f, we have $f(S \cup T') - f(S) \geq f(S' \cup T) - f(S')$.

$D_t \setminus D_\tau \subseteq B_t \setminus B_\tau$

$f_v(B_t) - f_v(B_\tau) \geq f_v(D_t) - f_v(D_\tau)$

$S = B_\tau, S' = D_\tau, T' = B_t \setminus B_\tau, T = D_t \setminus D_\tau$

$f_v(D_t) \geq f_v(D_\tau) + 1 - \theta_v$

$\Rightarrow f_v(B_t) \geq f_v(B_\tau) + 1 - \theta_v$

$D_{t+1} \setminus D_\tau \subseteq B_{t+1} \setminus B_\tau$

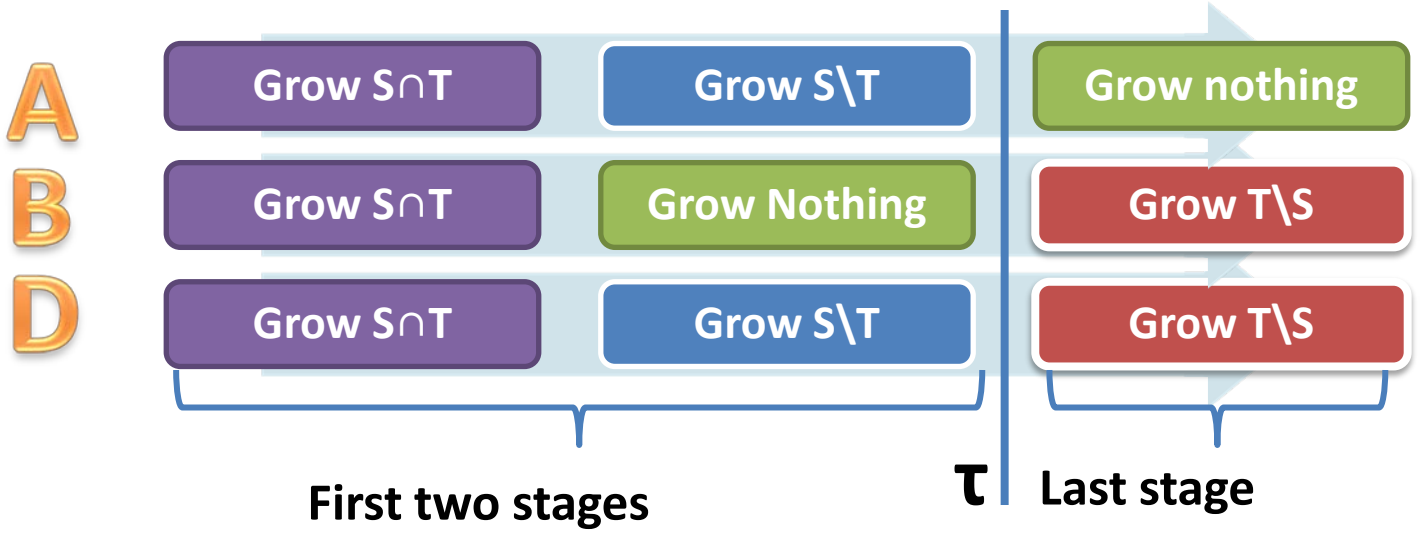
# Coupling: Wrapup

- Therefore we have:

$$D_t \setminus D_\tau \subseteq B_t \setminus B_\tau \subseteq B_t$$

$$A_t = D_\tau \text{ for all } t \text{ in the final stage}$$

$$D_t \subseteq A_t \cup B_t \,, C_t \subseteq A_t \cap B_t \text{ (Previously proved)}$$



# Further Generalization

- We have defined $\sigma(S)=E[|P_{end}||S]$.
- We can introduce a set function $\omega(\cdot)$ on $P_{end}$ and define the influence spread as $\sigma_\omega(S)=E[\omega(P_{end})|S]$ instead.

**Theorem:** Under the general threshold model with monotone and submodular $f_v$ and $\omega$, $\sigma_\omega(S)$ is monotone and submodular. [Mossel/Roch 2007]

# Further Generalization: Proof

Assume $C_{end} \subseteq A_{end} \cap B_{end}$ and $D_{end} \subseteq A_{end} \cup B_{end}$.

Then $\omega(A_{end}) + \omega(B_{end})$

$$\geq \omega(A_{end} \cap B_{end}) + \omega(A_{end} \cup B_{end})$$

$$\geq \omega(C_{end}) + \omega(D_{end}).$$

Taking expectation, we have

$$\sigma_\omega(S) + \sigma_\omega(T) \geq \sigma_\omega(S \cap T) + \sigma_\omega(S \cup T).$$

# Conclusion

- General Threshold Model generalizes many popular diffusion models.

**Theorem:** Under the general threshold model with monotone and submodular $f_v$ and $\omega$, $\sigma_\omega(S)$ is monotone and submodular. [Mossel/Roch 2007]

- Proof methodology: Coupling (piecemeal growth & antisense coupling)

# Algorithm for Influence Maximization

**Corollary:** The greedy algorithm is a (1-1/e) approximation to solve the influence maximization problem under general threshold model.

Algorithm 1: Greedy(k)

1: initialize S to empty set

2: for i = 1 to k do

3: select $u = \arg\max_{v \in V \setminus S} (\sigma(S \cup \{v\}) - \sigma(S))$

4: $S = S \cup \{u\}$

5: end for

6: return $S$

# Algorithm for Influence Maximization

Algorithm 1: Greedy(k)

1: initialize S to empty set

2: for i = 1 to k do

3: select $u = \arg\max_{v \in V \setminus S} (\sigma(S \cup \{v\}) - \sigma(S))$

4: $S = S \cup \{u\}$

5: end for

6: return $S$

- Time complexity: O(knCm)
- Where n=|V|, m=|E|, **C** the times of Monte-Carlo simulation.

# Algorithm for Influence Maximization

| Name | Main Idea | Model | Guarantee | Reference |
|---|---|---|---|---|
| CELF | Lazy Forward optimization | All | 1-1/e | Leskovec et al. 2007 |
| CELF++ | Further optimization of CELF | All | 1-1/e | Goyal et al. 2011 |
| PMIA | Use directed tree structure | IC | No | Chen et al. 2010 |
| LDAG | Use DAG structure | LT | No | Chen et al. 2010 |
| IRIE | Use PageRank to initialize and update locally | IC | No | Chen et al. 2012 |
| CGA | Use community structure | IC | $1-e^{-\frac{1}{1+\Delta d\theta}}$ | Wang et al. 2010 |
| MSA | Simulated Annealing | All | No | Jiang et al. 2011 |

# Open Questions

- Different classes of activation function $f_v$ .
  - Local subadditive set function → Global subadditive influence spread σ(S)?

- Find approximation algorithm for solving the influence maximization problem under diffusion models with non-submodular influence spread σ(S).

# Thank you