# Controlling diffusion processes on networks

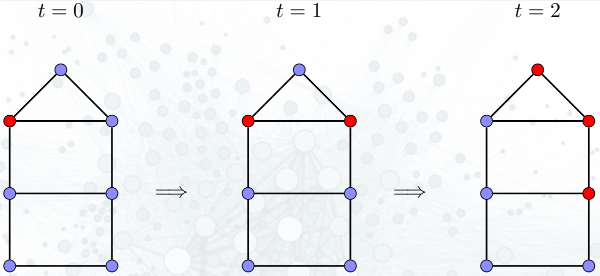
Anil Vullikanti
Department of Computer Science
Biocomplexity Institute & Initiative
University of Virginia

November 10, 2020

# Outline for lecture

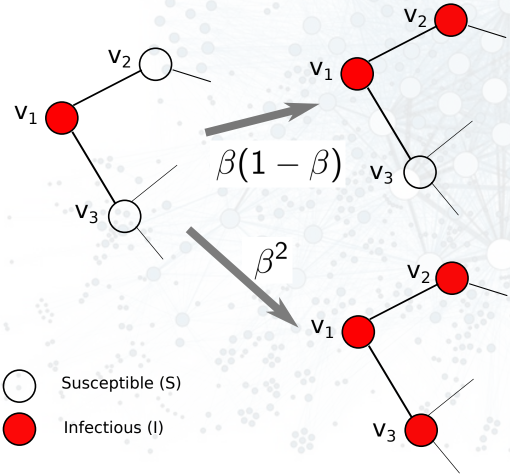
- Models of epidemic spread
- Maximizing diffusion
- Minimizing diffusion
- Summary

# Diffusion on networks



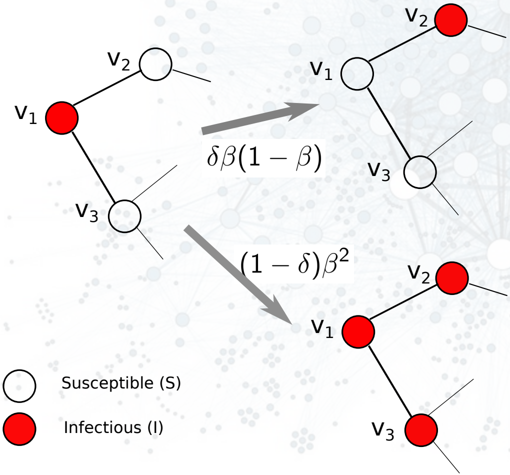

- Nodes in state 0 (inactive or uninfected) or 1 (active or infected)
- Switch state from 0 to 1, depending on neighbors
- Initially: set of seed nodes infected
- Large number of models, depending on the domain being modeled
  - Viral marketing: active node $\Rightarrow$ adopts a product. Goal: maximize number of active nodes
  - Spread of diseases: active node $\Rightarrow$ infected. Goal: minimize infections
  - Other phenomena: spread of innovations, ideologies, failures

# Stochastic model of diffusion on a network



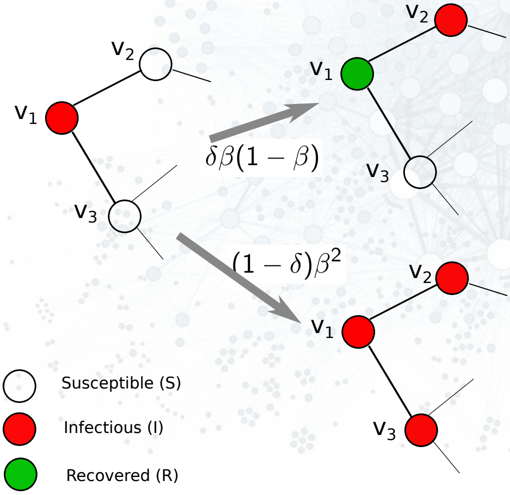

- Infectious node spreads infection to each neighbor independently with probability $\beta$ in each time step
- What happens to the infectious node in that time step
  - Nothing: remains infected (SI model)

# Stochastic model of epidemic spread on a network




- Infectious node spreads infection to each neighbor independently with probability $\beta$ in each time step
- What happens to the infectious node in that time step
  - Nothing: remains infected (SI model)
  - Becomes susceptible with probability $\delta$ (SIS model)

# Stochastic model of epidemic spread on a network




- Infectious node spreads infection to each neighbor independently with probability $\beta$ in each time step
- What happens to the infectious node in that time step
  - Nothing: remains infected (SI model)
  - Becomes susceptible with probability $\delta$ (SIS model)
  - Recovers with probability $\delta$, and never gets reinfected (SIR model)
  - Independent cascades (IC) model: special case of SIR, in which $\delta = 1$ (node recovers after 1 time step) ideologies

# Many other models of diffusion on networks

- Voter models: each node picks state of a random neighbor in each time step
- Threshold type models (complex contagion)
  - Each node $v$ has a threshold $\theta_v \in [0, 1]$
  - Node states are 0 or 1
  - Node $v$ switches from 0 to 1 if $\theta_v$ fraction of neighbors are 1
- Bi-threshold models: different thresholds for switching from 0 to 1 and from 1 to 0
- ......

# Influence maximization problem

- Consider IC model on a graph $G = (V, E)$, with probability $p$ of diffusion spread on an edge
- For $S \subset V$, let $F(S)$ denote the expected number of nodes which become active, if $S$ is initially active

**Problem**

Given $G$, $p$ and a budget $k$, find $S \subset V$ such that

- $|S| \leq k$
- $F(S)$ is maximized

# Submodularity

Function $F : 2^V \to \mathbb{R}$ is said to be a submodular set function if

$$\text{For all } S \subset T \subset V, v \notin T\text{: } F(S \cup \{v\}) - F(S) \geq F(T \cup \{v\}) - F(T)$$

- Captures diminishing marginal benefit property
- Greedy algorithm works well

# Greedy algorithm for Influence Maximization problem

**Main idea**

Iteratively add node to $S$ which gives maximum increase in $F(S)$

# Greedy algorithm for Influence Maximization problem

**Main idea**

Iteratively add node to $S$ which gives maximum increase in $F(S)$

- Initialize $S \leftarrow \varnothing$
- Repeat while $|S| < k$:
  - Pick the $v \in V - S$ that maximizes $F(S \cup \{v\}) - F(S)$
  - $S \leftarrow S \cup \{v\}$

# How do we compute $F(S)$?

- Exact computation of $F(S)$ is #P-hard
- Monte-Carlo sampling gives a good approximation

# Approximating $F(S)$: Monte-Carlo (MC) sampling

**Given**:

- Graph $G$, set $S$, parameters $\epsilon, \delta$
- Algorithm $\mathcal{A}$ which does a simulation of the IC model. Let $\mathcal{A}(S)$ denote the number of nodes activated in a stochastic simulation

**for** $i = 1$ to $T = \frac{n^2}{\epsilon^2}\ln(2/\delta)$ **do**
    $X_i = \mathcal{A}(S)$
**end for**
Let $X = \sum_{i=1}^{T} X_i$
Return $X/T$

**Lemma**

$X/T \in [(1-\epsilon)F(S), (1+\epsilon)F(S)]$, *with probability at least* $1-\delta$.

# Analysis of MC sampler for computing $F(S)$

$$
\begin{aligned}
E[X] &= \sum_i E[X_i] \text{ (Linearity of expectation)} \\
&= TF(S) \text{ since } E[X_i] = F(S) \\
\Rightarrow E[X/T] &= F(S)
\end{aligned}
$$

However, $X/T$ is a random variable. How do we argue that it is close to $F(S)$?

# Analysis of MC sampler for computing $F(S)$

Let $Y_i = X_i/n$. $E[Y_i] = F(S)/n \in (0,1]$

**Theorem (Hoeffding's bound)**

*Let $Y = \sum_{i=1}^{T} Y_i$, and the $Y_i$'s are independent random variables with $Y_i \in [0,1]$. Let $\gamma \in (0,1)$. Then,*

$$\Pr[|Y - E[Y]| > \gamma T] \leq 2e^{-2T\gamma^2}$$

Can be rewritten as: $\Pr[|Y/T - E[Y/T]| > \gamma] \leq 2e^{-2T\gamma^2}$

# Analysis of MC sampler for computing $F(S)$

- $Y_i = X_i/n$. $E[Y_i] = F(S)/n \in (0,1]$
- Apply Hoeffding's bound to $Y = \sum_{i=1}^{T} Y_i$, since $Y_i$'s are independent. Choose $\gamma = \epsilon F(S)/n \geq \epsilon/n$ as $F(S) \geq 1$
- For $T = \frac{n^2}{\epsilon^2}\ln(2/\delta)$: $T\gamma^2 = \frac{n^2}{\epsilon^2}\ln(2/\delta)\frac{\epsilon^2 F(S)^2}{n^2} \geq \ln(2/\delta)$
- $\frac{Y}{T} = \frac{1}{T}\frac{\sum_i X_i}{n} = \frac{X}{nT}$
- $E[Y/T] = \frac{TF(S)}{nT} = \frac{F(S)}{n}$
- $E[Y/T] - \gamma = \frac{F(S)}{n} - \epsilon\frac{F(S)}{n} = (1-\epsilon)\frac{F(S)}{n}$
- $E[Y/T] + \gamma = \frac{F(S)}{n} + \epsilon\frac{F(S)}{n} = (1+\epsilon)\frac{F(S)}{n}$

$$
\begin{aligned}
\Pr\left[\left|\frac{Y}{T} - E\left[\frac{Y}{T}\right]\right| > \gamma\right] &= \Pr\left[\frac{Y}{T} < (1-\epsilon)\frac{F(S)}{n} \text{ or } \frac{Y}{T} > (1+\epsilon)\frac{F(S)}{n}\right] \\
&\leq 2e^{-2T\gamma^2} \\
&\leq 2e^{-ln(2/\delta)} \\
&= \delta
\end{aligned}
$$

# Analysis of MC sampler for computing $F(S)$

$$
\begin{aligned}
\Pr\left[\frac{Y}{T} < (1-\epsilon)\frac{F(S)}{n} \text{ or } \frac{Y}{T} > (1+\epsilon)\frac{F(S)}{n}\right] &\leq \delta \\
\Rightarrow \Pr\left[\frac{X}{nT} < (1-\epsilon)\frac{F(S)}{n} \text{ or } \frac{X}{nT} > (1+\epsilon)\frac{F(S)}{n}\right] &\leq \delta \\
\Rightarrow \Pr\left[\frac{X}{T} < (1-\epsilon)F(S) \text{ or } \frac{X}{T} > (1+\epsilon)F(S)\right] &\leq \delta \\
\Rightarrow \Pr\left[\frac{X}{T} \in [(1-\epsilon)F(S), (1+\epsilon)F(S)]\right] &\geq 1-\delta
\end{aligned}
$$

# Analysis of greedy algorithm

**Lemma**

*If $F(\cdot)$ is a monotone, non-negative submodular function, then $F(S) \geq (1 - 1/e)F(S^*)$, where $S^* = argmax_{T:|T| \leq k} F(T)$ is an optimal solution*

- Let the greedy algorithm pick nodes $v_1, \ldots, v_k$.
- Let $S_i = \{v_1, \ldots, v_i\}$. Let $S_0 = \emptyset$, $F(S_0) = 0$
- Let $W_i = S^* \cup S_i$
- Let $\delta_i = F(S_i) - F(S_{i-1})$
- Claim: $F(W_i) \leq F(S_i) + k\delta_{i+1}$ (by submodularity)
- $F(S^*) \leq F(W_i) \leq F(S_i) + k\delta_{i+1}$ (monotonicity)
- $\delta_{i+1} \geq \frac{1}{k}(F(S^*) - F(S_i)) \Rightarrow F(S_{i+1}) \geq F(S_i) - \frac{1}{k}(F(S^*) - F(S_i))$
- Claim: $F(S_i) \geq (1 - (1 - \frac{1}{k})^i)F(S^*)$ for all $i$

# Proof (continued)

**Claim**

$F(W_i) \leq F(S_i) + k\delta_{i+1}$

- Let $S^* = \{u_1, \ldots, u_k\}$ (assume $S^* \cap S_i = \emptyset$)
- For all $j \leq i$: $F(S_i \cup \{u_1, \ldots, u_{j+1}\}) - F(S_i \cup \{u_1, \ldots, u_j\}) \leq F(S_i \cup \{u_{j+1}\}) - F(S_i)$ (by submodularity)
- $F(S_i \cup \{u_{j+1}\}) - F(S_i) \leq F(S_i \cup \{v_{j+1}\}) - F(S_i) = \delta_{i+1}$ (greedy choice)
- Summing over all $j$: $F(S_i \cup \{u_1, \ldots, u_k\}) - F(S_i) \leq k\delta_{i+1}$

# Proof (continued)

**Claim**

$F(S_i) \geq (1 - (1 - \frac{1}{k})^i)F(S^*)$ for all $i$

- Assuming Claim, we have $F(S_k) \geq (1 - (1 - \frac{1}{k})^k)F(S^*)$
- $1 - 1/k \leq e^{-1/k} \Rightarrow (1 - 1/k)^k \leq e^{-1}$
- $\Rightarrow (1 - (1 - \frac{1}{k})^k) \geq (1 - 1/e)$
- $\Rightarrow F(S_k) \geq (1 - 1/e)F(S^*)$

# Inductive proof of Claim

**Claim**

$F(S_i) \geq (1 - (1 - \frac{1}{k})^i)F(S^*)$ for all $i$

- Base case: $i = 0 \Rightarrow F(S_0) \geq 0$
- Assume inductive hypothesis holds for $i$

$$
\begin{aligned}
F(S_{i+1}) &\geq F(S_i) + \frac{1}{k}(F(S^*) - F(S_i)) \\
&\geq F(S_i)(1 - 1/k) + \frac{1}{k}F(S^*) \\
&\geq (1 - 1/k)(1 - (1 - \frac{1}{k})^i)F(S^*) + \frac{1}{k}F(S^*) \\
&\geq F(S^*)(1 - 1/k - (1 - \frac{1}{k})^{i+1} + 1/k) \\
&= F(S^*)(1 - (1 - \frac{1}{k})^{i+1})
\end{aligned}
$$