Probabilistic Contagion and Models of Influence
In decision-based models nodes make decisions based on pay-off benefits of adopting one strategy or the other.

In epidemic spreading:
- Lack of decision making
- Process of contagion is complex and unobservable
  - In some cases it involves (or can be modeled as) randomness
Example with $k=3$

High contagion probability: The disease spreads

Low contagion probability: The disease dies out
**Epidemic Model based on Random Trees**

- (a variant of branching processes)
- A patient meets \(d\) new people
- With probability \(q > 0\) she infects each of them

**Q:** For which values of \(d\) and \(q\) does the epidemic run forever?

**Run forever:**

\[
\lim_{h \to \infty} P\left[ \text{a node at depth } h \text{ is infected} \right] > 0
\]

**Die out:**

\[
\lim_{h \to \infty} P\left[ \text{a node at depth } h \text{ is infected} \right] = 0
\]
• $p_h = \text{prob. a node at depth } h \text{ is infected}$

• **We need:** $\lim_{h \to \infty} p_h = ?$ (based on $q$ and $d$)
  
  ▪ We are reasoning about a behavior at the root of the tree. Once we get a level out, we are left with identical problem of depth $h - 1$.

• **Need recurrence for $p_h$**

\[
p_h = 1 - (1 - q \cdot p_{h-1})^d
\]

No infected node at depth $h$ from the root

• $\lim_{h \to \infty} p_h = \text{result of iterating}$

\[
f(x) = 1 - (1 - q \cdot x)^d
\]

 ▪ Starting at the root: $x = 1$ (since $p_1 = 1$)

We iterate:

$x_1 = f(1)$

$x_2 = f(x_1)$

$x_3 = f(x_2)$
If we want the epidemic to die out, then iterating $f(x)$ must go to zero. So, $f(x)$ must be below $y = x$.

- What’s the shape of $f(x)$?
**Fixed Point:** \( f(x) = 1 - (1 - qx)^d \)

What do we know about the shape of \( f(x) \)?

- \( f(0) = 0 \)
- \( f(1) = 1 - (1 - q)^d < 1 \)
- \( f'(x) = q \cdot d(1 - qx)^{d-1} \)
- \( f'(0) = q \cdot d \)

\( f'(x) \) is **monotone** non-increasing on \([0,1]\)!

- **\( f'(x) \) is monotone:** If \( g'(y) > 0 \) for all \( y \) then \( g(y) \) is monotone.
  - In our case, \( 0 \leq x, q \leq 1, d > 1 \) so \( f'(x) > 0 \), so \( f(x) \) is monotone.

- **\( f'(x) \) non-increasing:** since term \((1-qx)^{d-1}\) in \( f'(x) \) is decreasing as \( x \) decreases.

\( x \) … prob. a node at level \( h-1 \) is infected.

We start at \( x = 1 \) because \( p_1 = 1 \).

\( f(x) \) … prob. a node at level \( h \) is infected

\( q \) … infection prob.

\( d \) … degree

\( f(x) \) is \( y = x = 1 \)

Going to the first fixed point

\( y = f(x) \)
Fixed Point: When is this zero?

For the epidemic to die out we need $f(x)$ to be below $y = x$!

So: $f'(0) = q \cdot d < 1$

$$\lim_{h \to \infty} p_h = 0 \text{ when } q \cdot d < 1$$

$q \cdot d = \text{expected # of people that get infected}$

Reproductive number $R_0 = q \cdot d$: There is an epidemic if $R_0 \geq 1$
Reproductive number $R_0 = q \cdot d$:
- It determines if the disease will spread or die out.
- There is an epidemic if $R_0 \geq 1$

Only $R_0$ matters:
- $R_0 \geq 1$: epidemic never dies and the number of infected people increases exponentially
- $R_0 < 1$: Epidemic dies out exponentially quickly
Measures to Limit the Spreading

- When $R_0$ is close 1, slightly changing $q$ or $d$ can result in epidemics dying out or happening
  - Quarantining people/nodes [reducing $d$]
  - Encouraging better sanitary practices reduces germs spreading [reducing $q$]
- HIV has an $R_0$ between 2 and 5
- Measles has an $R_0$ between 12 and 18
- Ebola has an $R_0$ between 1.5 and 2
Application: Social cascades on Flickr and estimating $R_o$ from real data

[Characterizing social cascades in Flickr](http://example.com). Cha et al. ACM WOSN 2008
Flickr social network:
- Users are connected to other users via friend links
- A user can “like/favorite” a photo

Data:
- 100 days of photo likes
- Number of users: 2 million
- 34,734,221 likes on 11,267,320 photos
Cascades on Flickr

- Users can be exposed to a photo via social influence (cascade) or external links
- Did a particular like spread through social links?
  - No, if a user likes a photo and if none of his friends have previously liked the photo
  - Yes, if a user likes a photo after at least one of her friends liked the photo → Social cascade
- Example social cascade: A → B and A → C → E
How to estimate $R_0$ from real data?

- **Recall:** $R_0 = q \times d$

- **Estimate of $R_0$:**
  - Estimating $q$: Given an infected node count the proportion of its neighbors subsequently infected and average.
  - Then:
    $$R_0 = q \times d \times \frac{avg(d_i^2)}{(avg d_i)^2}$$

- **Empirical $R_0$:**
  - Given start node of a cascade, count the fraction of directly infected nodes and proclaim that to be $R_0$
$R_0$ correlation across all photos

- Data from top 1,000 photo cascades
- Each + is one cascade
Discussion

The basic reproduction number of popular photos is between 1 and 190.

This is much higher than very infectious diseases like measles, indicating that social networks are efficient transmission media and online content can be very infectious.
Epidemic models
Virus Propagation: 2 Parameters:

- **(Virus) Birth rate $\beta$:**
  - Probability that an infected neighbor attacks

- **(Virus) Death rate $\delta$:**
  - Probability that an infected node heals
More Generally: S+E+I+R Models

- **General scheme for epidemic models:**
  - Each node can go through phases:
    - Transition probs. are governed by the model parameters

```
S...susceptible
E...exposed
I...infected
R...recovered
Z...immune
```
SIR Model

- **SIR model**: Node goes through phases
  - Models chickenpox or plague:
    - Once you heal, you can never get infected again
  - **Assuming perfect mixing** (The network is a complete graph) the model dynamics are:

\[
\begin{align*}
\frac{dS}{dt} &= -\beta SI \\
\frac{dR}{dt} &= \delta I \\
\frac{dI}{dt} &= \beta SI - \delta I
\end{align*}
\]
**SIS Model**

- Susceptible-Infective-Susceptible (SIS) model
- Cured nodes immediately become susceptible
- Virus “strength”: $s = \beta / \delta$
- Node state transition diagram:

   - Infected by neighbor with prob. $\beta$
   - Cured with prob. $\delta$
### SIS Model

- **Models flu:**
  - Susceptible node becomes infected
  - The node then heals and become susceptible again

- **Assuming perfect mixing (a complete graph):**

\[
\frac{dS}{dt} = -\beta SI + \delta I
\]

\[
\frac{dI}{dt} = \beta SI - \delta I
\]
Question: Epidemic threshold $\tau$

- **SIS Model:**
  Epidemic threshold of an arbitrary graph $G$ is $\tau$, such that:
  - If virus “strength” $s = \beta / \delta < \tau$ the epidemic cannot happen (it eventually dies out)

- **Given a graph what is its epidemic threshold?**
Fact: We have no epidemic if:

\[ \frac{\beta}{\delta} < \tau = \frac{1}{\lambda_{1,A}} \]

(Virus) Death rate

(Virus) Birth rate

Epidemic threshold

\[ \lambda_{1,A} \] alone captures the property of the graph!
Experiments (AS graph)

Autonomous Systems Graph

Number of Infected Nodes vs. Time

- \( \delta = 0.05 \)
- \( \delta = 0.06 \)
- \( \delta = 0.07 \)

\( \beta = 0.001 \)

10,900 nodes and 31,180 edges

\[ s = \frac{\beta}{\delta} \]

- \( s > \tau \) (above threshold)
- \( s = \tau \) (at the threshold)
- \( s < \tau \) (below threshold)

[Wang et al. 2003]
Experiments

- Does it matter how many people are initially infected?

(a) Below the threshold, \( s = 0.912 \)

(b) At the threshold, \( s = 1.003 \)

(c) Above the threshold, \( s = 1.1 \)
Modeling Ebola with SEIR

[Gomes et al., Assessing the International Spreading Risk Associated with the 2014 West African Ebola Outbreak, PLOS Current Outbreaks, '14]
\textbf{Example: Ebola}

\begin{itemize}
  \item \textbf{S}: susceptible individuals,
  \item \textbf{E}: exposed individuals,
  \item \textbf{I}: infectious cases in the community,
  \item \textbf{H}: hospitalized cases,
  \item \textbf{F}: dead but not yet buried,
  \item \textbf{R}: individuals no longer transmitting the disease
\end{itemize}

\begin{table}
\begin{tabular}{|c|c|}
\hline
Transition & Transition rate \\
\hline
(S,E) \to (S-1, E+1) & (\beta_1 SI + \beta_H SH + \beta_F SF)/N \\
(E,I) \to (E-1, I+1) & \alpha E \\
(I,H) \to (I-1, H+1) & \gamma_h \theta_1 I \\
(H,F) \to (H-1, F+1) & \gamma_d h \delta_2 H \\
(F,R) \to (F-1, R+1) & \gamma_f F \\
(I,R) \to (I-1, R+1) & \gamma_i (1 - \theta_1)(1 - \delta_1) I \\
(I,F) \to (I-1, F+1) & \delta_1 (1 - \theta_1) \gamma_d I \\
(H,R) \to (H-1, R+1) & \gamma_i h (1 - \delta_2) H \\
\hline
\end{tabular}
\end{table}

\[\text{Gomes et al., Assessing the International Spreading Risk Associated with the 2014 West African Ebola Outbreak, PLOS Current Outbreaks, '14}\]
Example: Ebola, $R_0 = 1.5 - 2.0$

Read an article about how to estimate $R_0$ of ebola.

[Gomes et al., 2014]
Application: Rumor spread modeling using SEIZ model

References:
SEIZ model: Extension of SIS model

**Susceptible** S  Twitter accounts

**Infected** I  Believe news / rumor, (I) post a tweet

**Exposed** E  Be exposed but not yet believe

**Skeptics** Z  Skeptics, do not tweet

Disease  Twitter
Recap: SIS model

\[
\frac{d[S]}{dt} = \dot{S} = -\beta SI + \alpha I
\]
\[
\frac{d[I]}{dt} = \dot{I} = \beta SI - \alpha I
\]

\[S = S(t), I = I(t)\]
\[\beta = \text{rate of contact between 2 individuals}\]
\[\alpha = \text{rate of recovery}\]

Disease Applications:
- Influenza
- Common Cold

Twitter Application Reasoning:
- An individual either believes a rumor (I),
- or is susceptible to believing the rumor (S)
Details of the SEIZ model

Notation:

- $S$ = Susceptible
- $I$ = Infected
- $E$ = Exposed
- $Z$ = Skeptics

Mathematical equations:

\[
\begin{align*}
\frac{d[S]}{dt} &= -\beta S \frac{I}{N} - bS \frac{Z}{N} \\
\frac{d[E]}{dt} &= (1 - p)\beta S \frac{I}{N} + (1 - l)bS \frac{Z}{N} - \rho E \frac{I}{N} - \epsilon E \\
\frac{d[I]}{dt} &= p\beta S \frac{I}{N} + \rho E \frac{I}{N} + \epsilon E \\
\frac{d[Z]}{dt} &= lbS \frac{Z}{N}
\end{align*}
\]
Dataset

Tweets collected from eight stories: Four rumors and four real events

REAL EVENTS

- Boston Marathon Explosion. 04-15-2013
- Pope Resignation. 02-11-2013
- Venezuela’s refinery explosion. 08-25-2012
- Michelle Obama at the 2013 Oscars. 02-24-2013

RUMORS

- Obama injured. 04-23-2013
- Doomsday rumor. 12-21-2012
- Fidel Castro’s coming death. 10-15-2012
- Riots and shooting in Mexico. 09-05-2012
Method: Fitting SEIZ model to data

- SEIZ model is fit to each cascade to minimize the difference $|I(t) - \text{tweets}(t)|$:
  - $\text{tweets}(t) = \text{number of rumor tweets}$
  - $I(t) = \text{the estimated number of rumor tweets by the model}$
- Use grid-search and find the parameters with minimum error

```
SEIZ model
\downarrow
\begin{align*}
\text{Parameter} & \quad \text{Iteration} & \quad \text{minimize} & \quad |I(t) - \text{tweets}(t)| \\
\downarrow & & \quad \text{Solve ODE} & \quad \text{System} \\
\text{Optimal} & \quad \text{Parameter} & \quad \text{Set}
\end{align*}
```
Fitting to “Boston Marathon Bombing”

SEIZ model better models the real data, especially at initial points

\[
\text{Error} = \frac{\text{norm}(I - \text{tweets})}{\text{norm}(\text{tweets})}
\]
Fitting to “Pope resignation” data

SEIZ model better models the real data, especially at initial points.
Rumor detection with SEIZ model

By SEIZ model parameters

Notation:
S = Susceptible
I = Infected
E = Exposed
Z = Skeptics

New metric:

\[ R_{SI} = \frac{(1 - p) \beta + (1 - l)b}{\rho + \epsilon} \]

\( R_{SI} \), a kind of flux ratio, the ratio of effects entering E to those leaving E.
Rumor detection by $R_{SI}$

Parameters obtained by fitting SEIZ model efficiently identifies rumors vs. news
Independent Cascade Model
Initially some nodes $S$ are active
Each edge $(u,v)$ has probability (weight) $p_{uv}$

When node $u$ becomes active/infected:
- It activates each out-neighbor $v$ with prob. $p_{uv}$
- Activations spread through the network!
Independent Cascade Model

- **Independent cascade model is simple but requires many parameters!**
  - Estimating them from data is very hard
    - [Goyal et al. 2010]

- **Solution:** Make all edges have the same weight (which brings us back to the SIR model)
  - Simple, but too simple

- **Can we do something better?**
Exposures and Adoptions

- From exposures to adoptions
  - **Exposure**: Node’s neighbor exposes the node to the contagion
  - **Adoption**: The node acts on the contagion
Exposure Curves

- **Exposure curve:**
  - Probability of adopting new behavior depends on the total number of friends who have already adopted.

- **What’s the dependence?**

  - Prob. of adoption as a function of $k$ (number of friends adopting).

  - "Probabilistic" spreading: Viruses, Information

  - Critical mass: Decision making
Exposure Curves

- **From exposures to adoptions**
  - **Exposure**: Node’s neighbor exposes the node to information
  - **Adoption**: The node acts on the information

- **Examples of different adoption curves:**
  
  ![Exposure Curves Diagram](image)

  - **Probability of infection ever increases**
  - **Nodes build resistance**
Diffusion in Viral Marketing

- Senders and followers of recommendations receive discounts on products
  - 10% credit
  - 10% off

- Data: Incentivized Viral Marketing program
  - 16 million recommendations
  - 4 million people, 500k products

[Leskovec et al., TWEB ’07]
Exposure Curve: Validation

DVD recommendations
(8.2 million observations)
Group memberships spread over the network:

- **Red** circles represent existing group members
- **Yellow** squares may join

**Question:**

- How does prob. of joining a group depend on the number of friends already in the group?
Exposure Curve: LiveJournal

- LiveJournal group membership

![Graph showing the probability of joining LiveJournal groups as a function of the number of friends in the group.](image)
Twitter [Romero et al. ‘11]

- Aug ‘09 to Jan ’10, 3B tweets, 60M users

- Avg. exposure curve for the top 500 hashtags
- What are the most important aspects of the shape of exposure curves?
- Curve reaches peak fast, decreases after!
**Modeling the Shape of the Curve**

- **Persistence of** $P$ **is the ratio of the area under the curve** $P$ **and the area of the rectangle of height** $\max(P)$, **width** $\max(D(P))$
  - $D(P)$ **is the domain of** $P$
  - **Persistence measures the decay of exposure curves**

- **Stickiness of** $P$ **is** $\max(P)$
  - **Stickiness is the probability of usage at the most effective exposure**
Manually identify 8 broad categories with at least 20 HTs in each:

- Idioms and Music have lower persistence than that of a random subset of hashtags of the same size
- Politics and Sports have higher persistence than that of a random subset of hashtags of the same size
Exposure Curve: Stickiness

- Technology and Movies have lower stickiness than that of a random subset of hashtags
- Music has higher stickiness than that of a random subset of hashtags (of the same size)
Recap of this lecture

- Basic reproductive number $R_0$
- **General epidemic models**
  - SIR, SIS, SEIZ
  - Independent cascade model
  - Applications to rumor spread
  - Exposure curves