LIFE DATA EPIDEMIOLOGY

Lecture 8: Network spreading

Leonardo Badia

leonardo.badia@unipd.it

SI model over a network

- A message is spreading over a network
  - e.g. a gossip spreading by word of mouth
  - can we use the equations of the SI model?

$$\frac{ds}{dt} = -\beta sx, \quad \frac{dx}{dt} = \beta sx$$

- Model prediction was that the number of "infected" reaches 100% as a sigmoid (exponential at first, then flat): is it so?

Cascading as diffusion model

- Consider a general trend for imitation
- More than for epidemics, this is used for beneficial imitation (but same model) e.g.
  - **information acquisition**: you get some news and you spread it
  - **direct-benefit effects**: you imitate others with similar traits because of advantages (e.g., adopting compatible technology)

Cascading over networks

- We model direct-benefit relations over a network (seen as a graph) as follows:
  - nodes in the network can take either of two behaviors, called **A** and **B**
  - nodes get a utility (payoff) depending on what they do and their neighbors do
  - those playing **A** get $a > 0$ for each neighbor playing **A**, and 0 for those choosing **B**;
  - **B** gives either 0 or $b > 0$ per neighbor, resp.

Cascading over networks

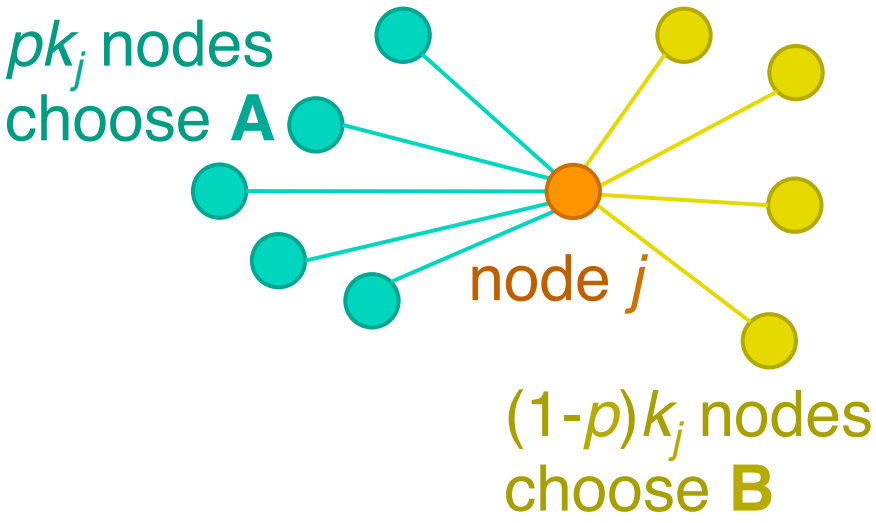
- The situation can be displayed as a table
  - in game theory this is a **coordination game** between any two nodes *i* and *j*

i \ *j*	A	B
A	***a***, ***a***	0, 0
B	0, 0	***b***, ***b***

  - in reality this is an n-person game whose payoff=**Σ**(local games)

Cascading over networks

- Each node selects its behavior in a **selfish** and **myopic** way
  - i.e., maximizing its own individual payoff and assuming that others do not change
    - *j* chooses depending on the neighbors
    - a fraction *p* (or 1-*p*) of them choose **A** (or **B**)

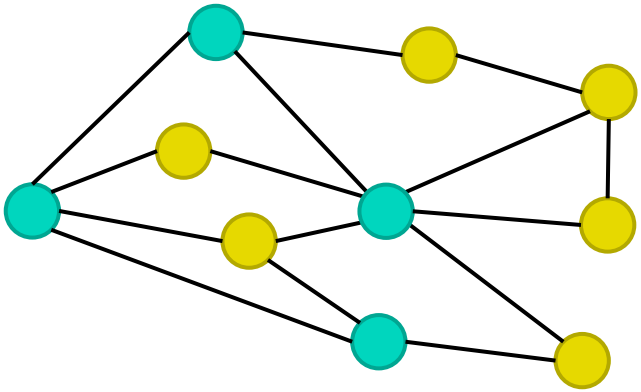


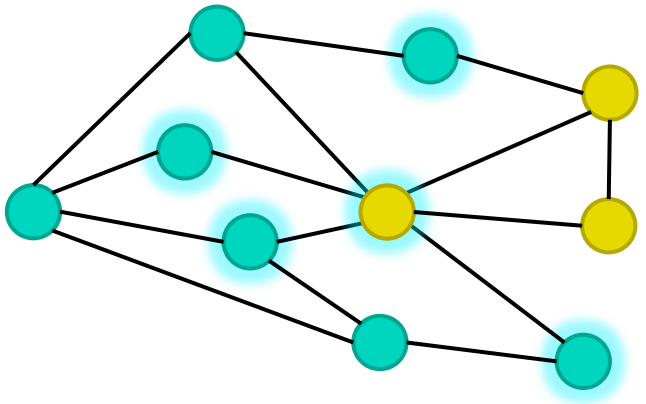
Cascading over networks

- The decision rule for node *j* is simple:
  - choose **A** if $k_j p\mathbf{a} > k_j(1-p)\mathbf{b}$, else choose **B**
  - this can be rearranged as $p > \boxed{\mathbf{b} / (\mathbf{a}+\mathbf{b})}$ call it p^*
  - in other words, node *j* follows a simple threshold rule: if $p > p^*$ choose **A**, else **B**
- In the end, a node follows what "most" of its neighbors do (but weighted on payoffs)

Equilibria of cascading

- In such a scenario we can think of a dynamic process as per this example
  - assume ***a***=5, ***b***=3 → the nodes highlighted do not like their current choice and change





Equilibria of cascading

- Iterating the process, we eventually reach
  - choice **A** for all nodes
  - choice **B** for all nodes
  - or some intermediate cases where both choices coexist (is this possible?)
- Especially, we may be interested in evaluating this for an innovation trend
  - e.g., $a > b$ although at t=0 most nodes adopt **B** and just a tiny fraction does **A**

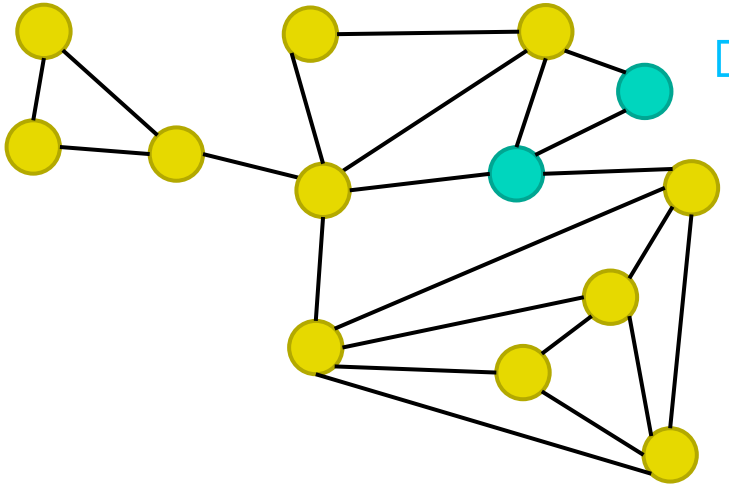
Dynamic model

- For the sake of simplicity, consider the "initial adopters" to always stick to **A**
  - we actually need an external motivation for these "pioneers of innovation" to choose **A** when everybody else does **B** instead
- It can be proven then that nodes can only switch **B**→**A** and not the opposite
  - due to the fact that the number of neighbors choosing innovation **A** can only increase

Complete cascade at p*

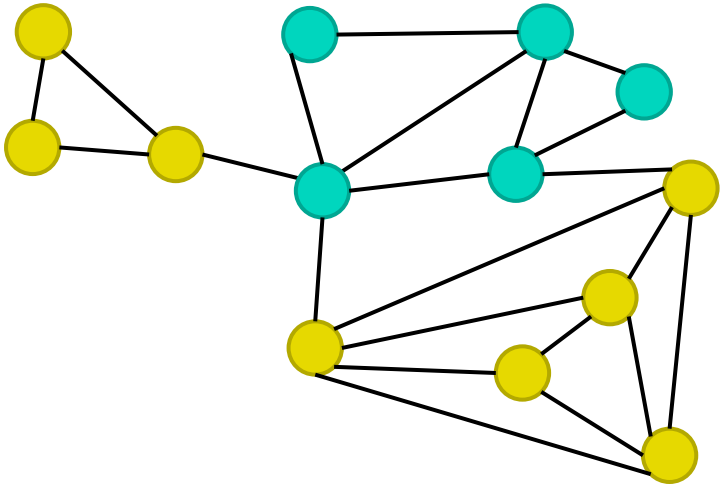
- The dynamic cascading makes some nodes to imitate the innovators → if this causes all nodes to eventually adopt A we say we have a **complete cascade**
  - this depends on *p**, the network topology, and where are the initial innovators

Complete cascade at p*



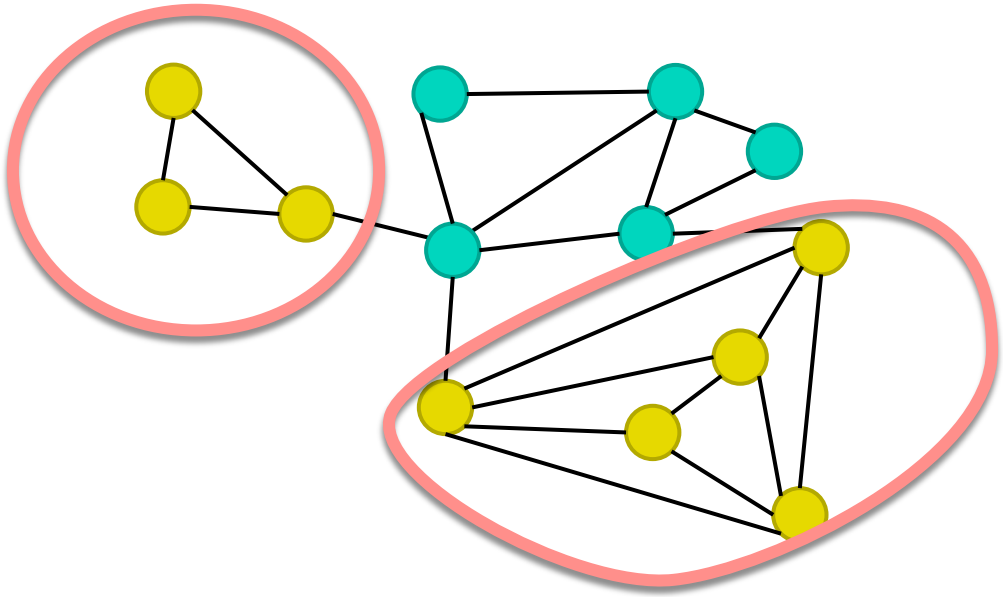
- The initial innovators

- After 3 steps
  (the process stops)



Complete cascade at p*

- What blocks the spreading over the entire network?



- The innovation process never gets past these tight-knit communities

Consequences on viral trends

- To disseminate content over an entire network, several aspects may be used
  - lowering threshold p* clearly helps
  - also, selecting the initial dissemination points as key nodes (well connected hubs)
  - finally: to have access to every community → more convenient to have just few connected disseminators in every region than a powerful dissemination in only one

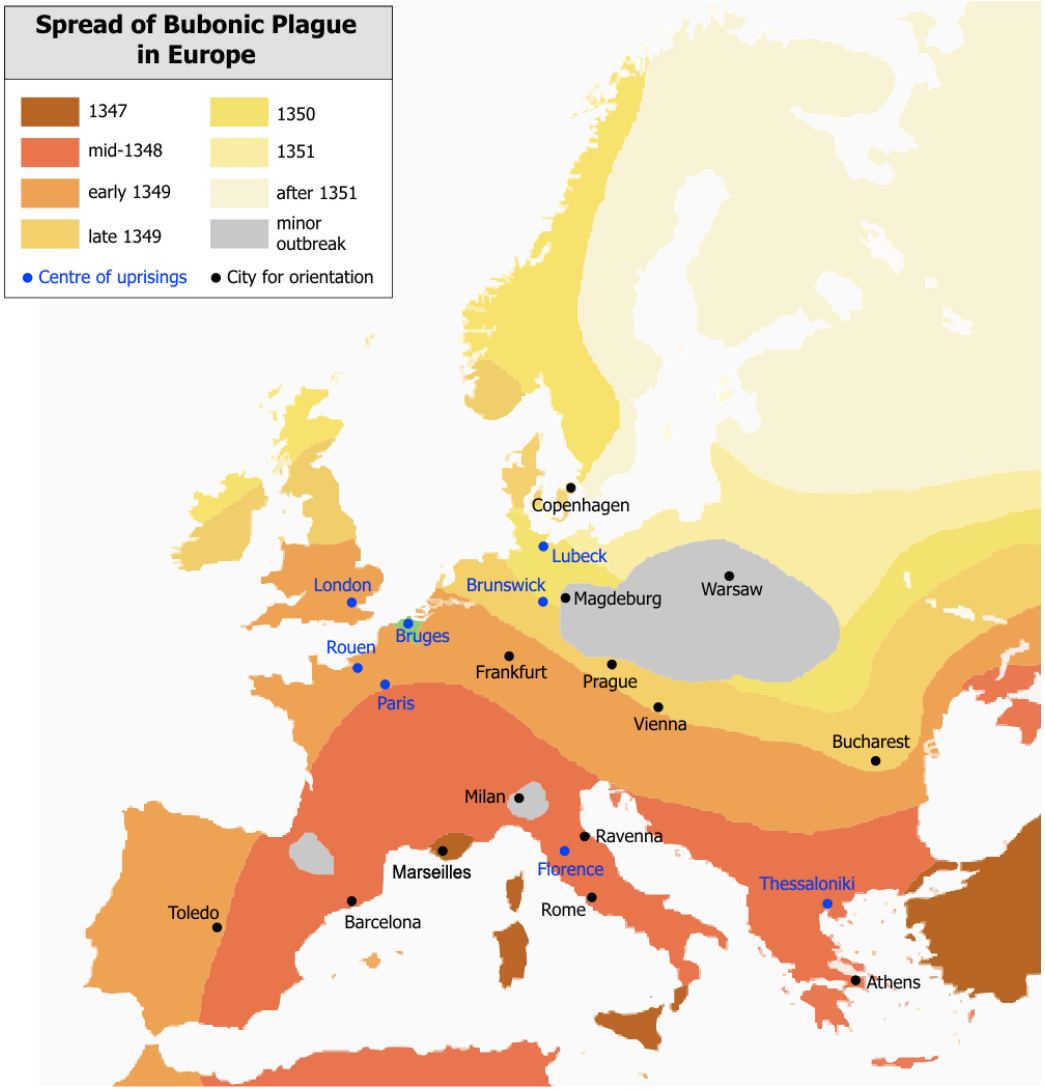
SI model over a network

- The SI model just use average values
  - $dx / dt = \beta s x$ means x grows **on average** proportional to s and x over a time unit
  - also we described β as the result of #contacts times (contagion I contact)
  - in a network with average degree $\langle k \rangle$, β is also proportional to $\langle k \rangle$, so we write it as $\beta \langle k \rangle$ (no longer the same β as before)
  - we can do it since we work on averages

SI model over a network

- This formulation can be implemented in every model seen so far (not only SI but also SIR, SIRS, SEIR, SICR...)
- Problem is, it only works under very unrealistic assumptions:
  - homogeneous mixing
  - all nodes have comparable degree
- Also, it does not tell us how the network structure influences the spreading

Bubonic plague in Europe



- Some regions are “spared” as they have a relatively low impact of the epidemics

Network epidemics

- A key idea [Vespignani, Pastor-Satorras] is to use a **degree-block approximation**
  - we consider all the nodes with same degree *k* as belonging to an ideal subnetwork
  - good modeling economy
- N_k = #{degree k nodes}, p_k = deg.distrib.
  X_k = #{infected nodes with degree k}
- Consider $x_k = X_k / N_k \rightarrow x = \Sigma_k x_k p_k$

Network epidemics

- Now, we rewrite the basic SI equation as:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \beta \langle k \rangle s x \quad \Rightarrow \quad \frac{\mathrm{d}x_k}{\mathrm{d}t} = \beta k (1 - x_k) \Theta_k$$

- We have k_{max} equations (one per degree)
- We keep the same rationale, but replace
  - $\langle k \rangle$ with the actual degree k
  - s_k with 1-x_k (so we only have one variable)
  - x_k with Θ_k = **density of infected neighbors** of a **susceptible** node with degree *k*

Early epidemic stages

- How to exploit $dx_k / dt = \beta k(1-x_k)\Theta_k$?
  - focus on small t (as in the standard SI)
  - this way, we can treat $x_k \approx 0$ in the r.h. side

→ We get $dx_k / dt = \beta k \Theta_k$

- We can show that in the absence of degree correlations, the term Θ_k is
  - indepedent of k (so we call it just Θ)
  - exponentially growing in t

Density of infected neighbors

- In the absence of degree correlation: probability that from an *h*-node, following a random link we reach a *k*-node

$$q_k = k\, p_k / \langle k \rangle$$

  independent of initial node's degree *h*
  - indeed, we reach a *k*-node since we have followed one of its *k* connections
  - and there is no correlation between *h* and *k*

Friendship paradox [Feld, 1961]

- My friends are more popular than me
  - This is actually true: nodes with high degrees are more likely to be counted as friends, and they skew the average of the no. of friends of friends
  - So $\langle k \rangle = \mathbb{E}[\#\text{friends}]$ is smaller than $\mathbb{E}[\#\text{friends of friends}]$
  - this is particularly visible on scale-free networks since they have hubs

Density of infected neighbors

- What is the probability that starting from a **susceptible** *h*-node, following a random link, we reach an **infected** *k*-node? It is

$$q'_k = (k\text{-}1)\, p_k / \langle k \rangle$$

still independent of initial node's degree h

  - not all *k* connections available: the *k*-node is infected; hence, one link points to another infected, from whom it got the disease, but we are starting from a susceptible!

Density of infected neighbors

- Hence $$\Theta_k = \Theta = \frac{\sum_k (k-1) p_k x_k}{\langle k \rangle}$$
- We can also relate the derivatives, which allows us to understand how Θ grows

$$\frac{d\Theta}{dt} = \sum_k \frac{(k-1) p_k}{\langle k \rangle} \frac{dx_k}{dt}$$

  - up to some modifications, this expression is true for not only SI, but also other models

back to SI equations

- Now, for SI we know $dx_k / dt = \beta k(1-x_k)\Theta_k$
  - if small $t \rightarrow$ $dx_k / dt = \beta k \Theta_k$
  - thus $$\frac{d\Theta}{dt} = \beta \sum_k \frac{(k^2 - k)p_k}{\langle k \rangle} \Theta = \beta \left(\frac{\langle k^2 \rangle}{\langle k \rangle} - 1 \right) \Theta$$
- Solving as usual we get $\Theta = \Theta_0 e^{Ft}$

  where $F = \beta(\langle k^2 \rangle - \langle k \rangle)/\langle k \rangle$, $\Theta_0 = x_0 (\langle k \rangle - 1)/\langle k \rangle$

choosing the x_0 initial infected with uniform random distribution

back to SI equations

- Inserting into the equations for x_k we get:

$$\frac{\mathrm{d}x_k}{\mathrm{d}t} = \beta k \; x_0 \left(\frac{\langle k \rangle - 1}{\langle k \rangle} \right) \mathrm{e}^{Ft}$$

that leads to $$x_k = x_0 \left(1 + \frac{k(\langle k \rangle - 1)}{\langle k^2 \rangle - \langle k \rangle} (\mathrm{e}^{Ft} - 1) \right)$$

  - thus, x_k depends on both degree distribution of the network and the specific value of k

back to SI equations

- We derived that $x_k = f(t) + k\, g(t)$ with both f and g being increasing functions
  - nodes with higher k are reached earlier
  - or, their "infected share" x_k grows faster
- $x = \Sigma_k\, x_k\, p_k \rightarrow$ $$x = x_0 \left(1 + \frac{\langle k \rangle^2 - \langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} (e^{Ft} - 1) \right)$$
  - especially, focus on $F = \beta(\langle k^2 \rangle - \langle k \rangle)/\langle k \rangle$

back to SI equations

- For an Erdős-Rényi model (random graph) we have $\langle k^2 \rangle = \langle k \rangle (1 + \langle k \rangle)$
  - we get $F = \beta \langle k \rangle$, $x = x_0 \left(1 + \frac{\langle k \rangle - 1}{\langle k \rangle} (e^{\beta \langle k \rangle t} - 1) \right)$
  - in other words, x grows exponentially with exponent $\beta \langle k \rangle$ (well, only for small t → after that it becomes a sigmoid)
- Fallback to SI (homogeneous mixing)!

SI over scale-free networks

- Scale-free with $\gamma > 3$ → $\langle k^2 \rangle$, $\langle k \rangle$ are finite
  - $F = \beta(\langle k^2 \rangle - \langle k \rangle) / \langle k \rangle$ is finite
  - similar behavior to E-R, i.e., exponential increase (although F can be bigger)
- Scale-free with $\gamma < 3$ → $\langle k^2 \rangle$ goes to ∞
  - contagion is very rapid (instantaneous)
  - this is because big hubs are both the first to be infected and also spread the disease very effectively (**super spreaders**)

Other networks

- We do not need a scale-free network to enhance the spreading of the disease
  - all that we require is that $\langle k^2 \rangle$ is large, so that there are big hubs acting as spreaders
- On the other hand, on regular networks (e.g. lattices) we see a reduced term F
  - consistent with the intuition that any node can infect at most a limited number

SIR model over networks

- SI → eventually everyone is infected
- What about SIR? We need to add a term:

$$dx_k / dt = \beta k(1-x_k-r_k)\Theta - \mu x_k$$

  - where again $\Theta = \dfrac{\sum_k (k-1)p_k x_k}{\langle k \rangle}$
  - and we set $(1-x_k-r_k) \approx 1$ so as to obtain

$$dx_k / dt = \beta k\Theta - \mu x_k$$

SIR model over networks

- Analogously to what we derived for SI:

$$d\Theta / dt = [\, \beta\, (\langle k^2 \rangle - \langle k \rangle)/\langle k \rangle - \mu]\, \Theta$$

- This leads us again to $\Theta = \Theta_0\, e^{Ft}$
  - but now $F = [\beta \langle k^2 \rangle - (\beta+\mu)\langle k \rangle]/\langle k \rangle$
  - i.e., the disease is not guaranteed to spread
  - F must be positive or else Θ vanishes: threshold criterion similar to the R_0 criterion

$F > 0$ if $\beta/\mu > \langle k \rangle / (\langle k^2 \rangle - \langle k \rangle)$

called spreading rate α

SIS model over networks

- We can write similar equations for SIS
- That is, we start from $\mathrm{d}x_k / \mathrm{d}t = \beta k \Theta - \mu x_k$
  - but now $\Theta = \Sigma_k \, k \, x_k \, p_k / \langle k \rangle$ as nodes can be susceptible even after being infective
  - thus, all neighbors are potential infectives
- We still get $\Theta = \Theta_0 \, e^{Ft}$, $F = [\beta \langle k^2 \rangle - \mu \langle k \rangle] / \langle k \rangle$
  - threshold criterion is now:

$F > 0$ if $\alpha = \beta / \mu > \langle k \rangle / \langle k^2 \rangle$

Threshold criterion

- Rules $\alpha^{(SIR)} > \langle k \rangle / (\langle k^2 \rangle - \langle k \rangle)$, $\alpha^{(SIS)} > \langle k \rangle / \langle k^2 \rangle$ relate "biological parameters" ($\alpha = \beta/\mu$) with the network topology and deg.distrib.
- We get a threshold α_C for contagion
  - threshold for SIS is slightly lower than that of SIR since we allow hubs to be infected and spreading the disease once again
  - note that threshold is only necessary but not sufficient (strictly speaking)

Threshold criterion

- On a random graph:

$$\alpha_C^{(SIR)} = 1/\langle k \rangle, \qquad \alpha_C^{(SIS)} = 1/(1+\langle k \rangle)$$

  - in both cases, the higher $\langle k \rangle$, the lower the threshold (but always finite) → easier spread

- But for a scale-free network, $\langle k^2 \rangle$ is bigger; for $\gamma < 3$ → $\langle k^2 \rangle \longrightarrow \infty$, i.e., thresholds=0
  - consequence: even a weak infection can still spread (by just reaching the right hub)

Threshold criterion

- Important difference with homogeneous models: on a scale-free network with $\gamma < 3$
  - it is **very easy** to spread a virus, as even contagions with low α can spread
  - the contagion is almost instantaneous as the exponential parameter is very high

→ all these conclusions are derived under the block-degree approximation but are valid also under more precise models