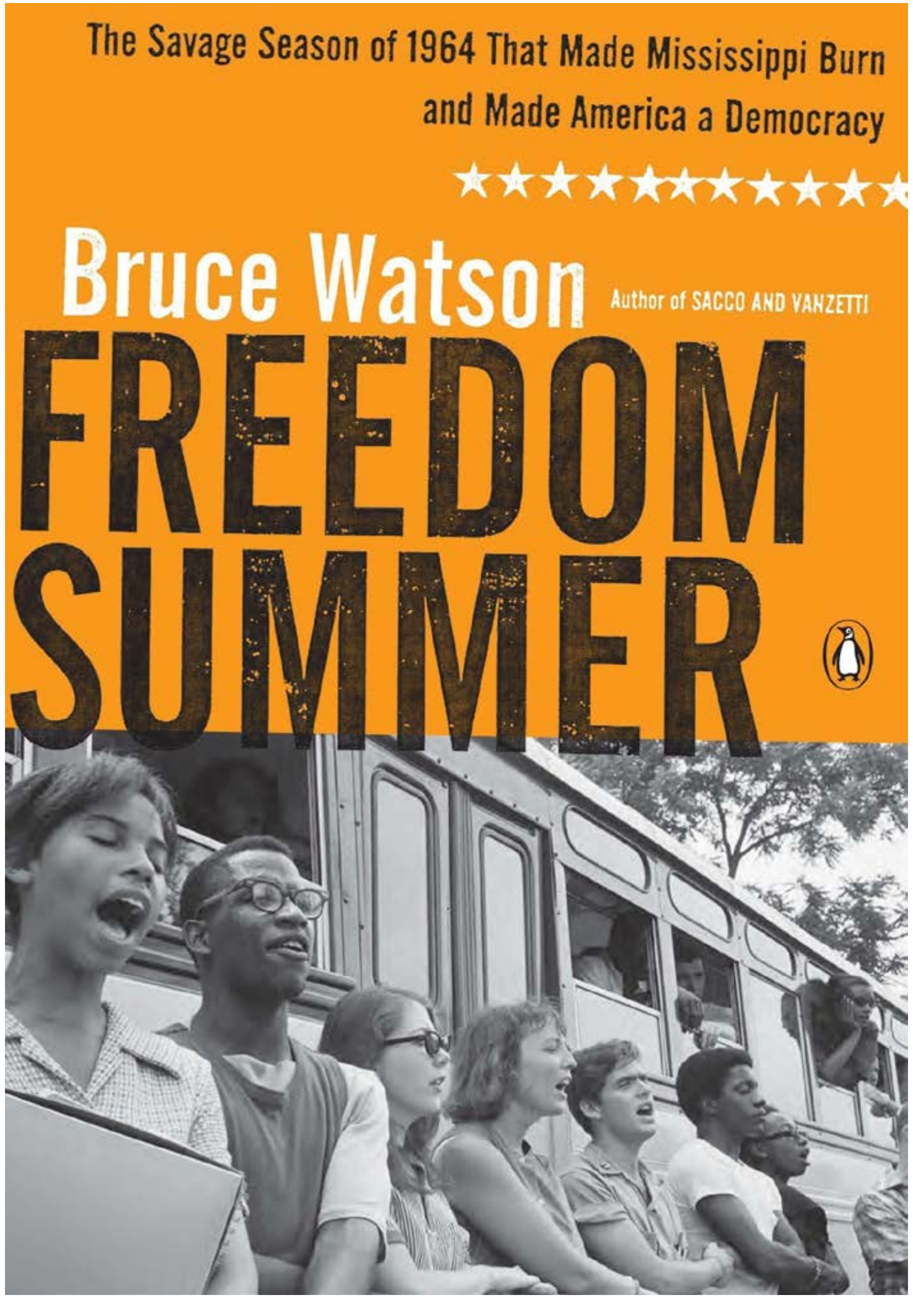
Diffusion in Networks

Vincent Buskens
Jiamin Ou

Summer School: Network Science
Day 5, Afternoon Session
2023



Memberships of civil rights activists in 1964- "Freedom Summer" : Despite the fact that weak ties offered an obvious advantage for rapidly increasing the movement's exposure, memberships **grew primarily through recruitment networks composed of strong ties**.

(Lisa Berkman and Ichiro Kawachi, 2011)

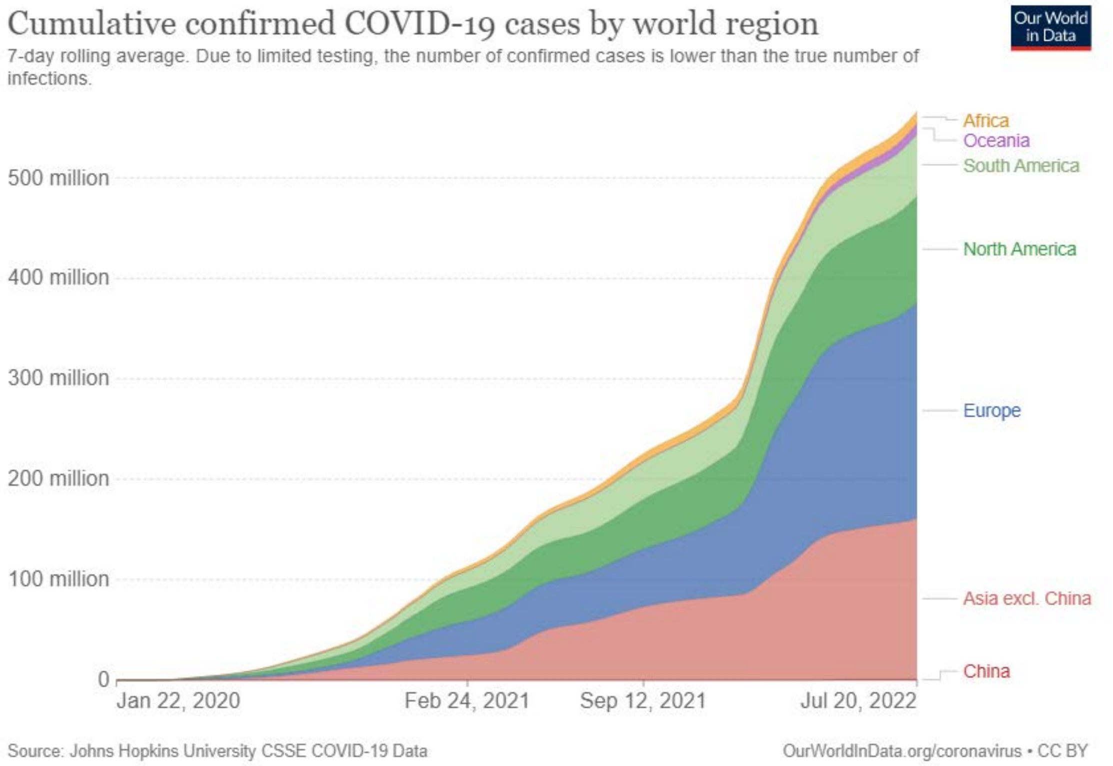
The spread of personal hygiene behaviour is much slower than COVID. Why?

Spread of COVID

VS

Spread of hand-washing, mask-wearing and other behaviours

567M by Jul 2022



Let’s try to find out:

Does social change spread in the network the same way as viruses and information do?

As a network scientist, what can you do?

Today's programe

- **Simple contagion**

  Mechanism and the strength of 'weak ties'

- **Diffusion model for simple contagion**

  Independent cascade model and other variants

- **Complex contagion**

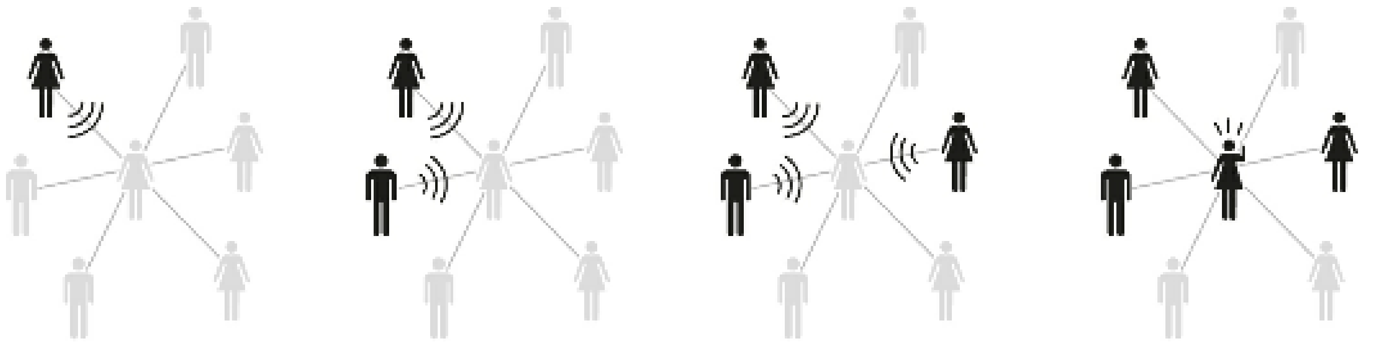
  Mechanism and the strength of 'strong ties'

- **Diffusion model for complex contagion**

  Threshold model and seeding

Mechanism of complex contagion

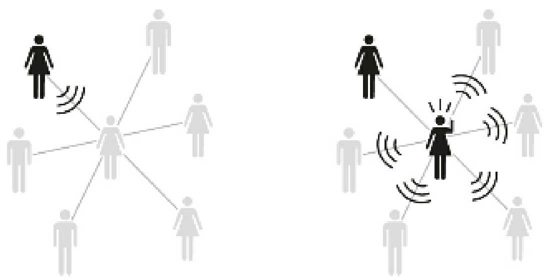
- Multiple (and credible) sources are required for transmission



Examples

- Change of diet
- High-risk social movements
- Other risky/costly behaviours

Recall for simple contagion:

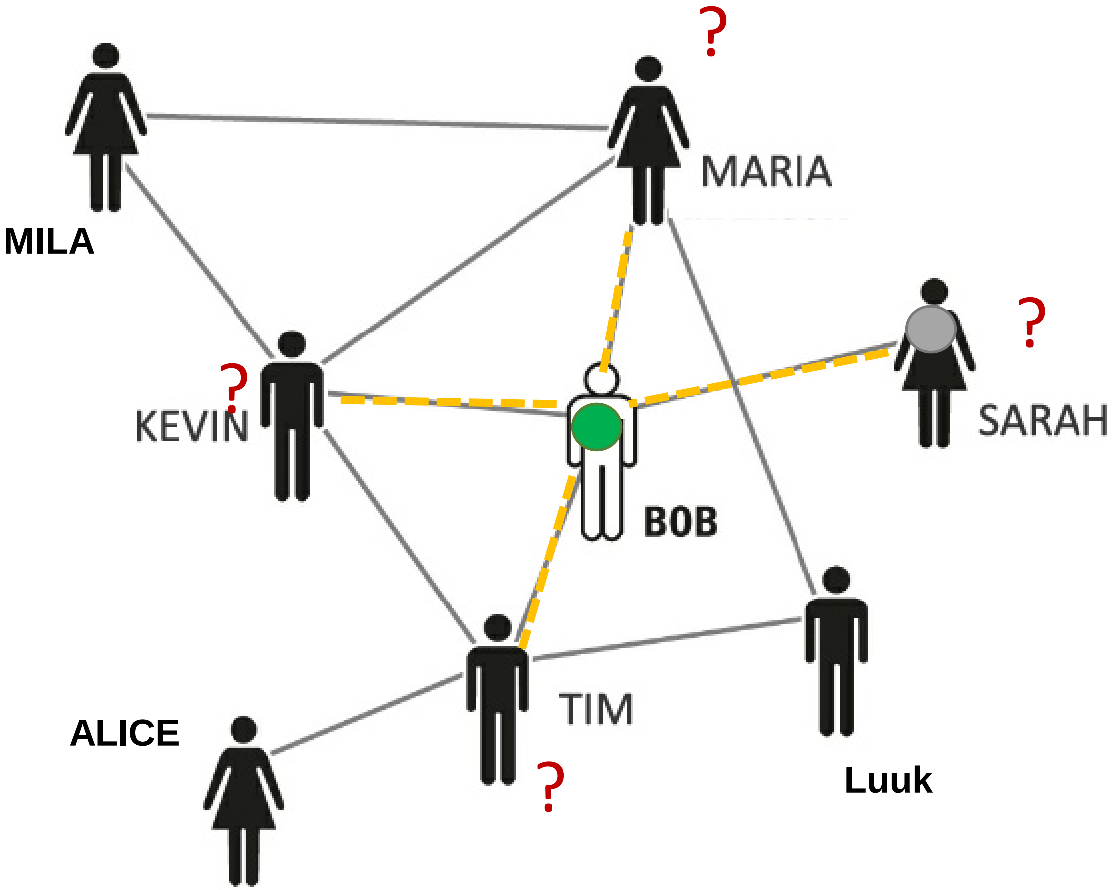


Demonstration of complex contagion in a small network

Bob is now considering joining an illegal climate change protest

(High-risk social movement)

1st round: Bob

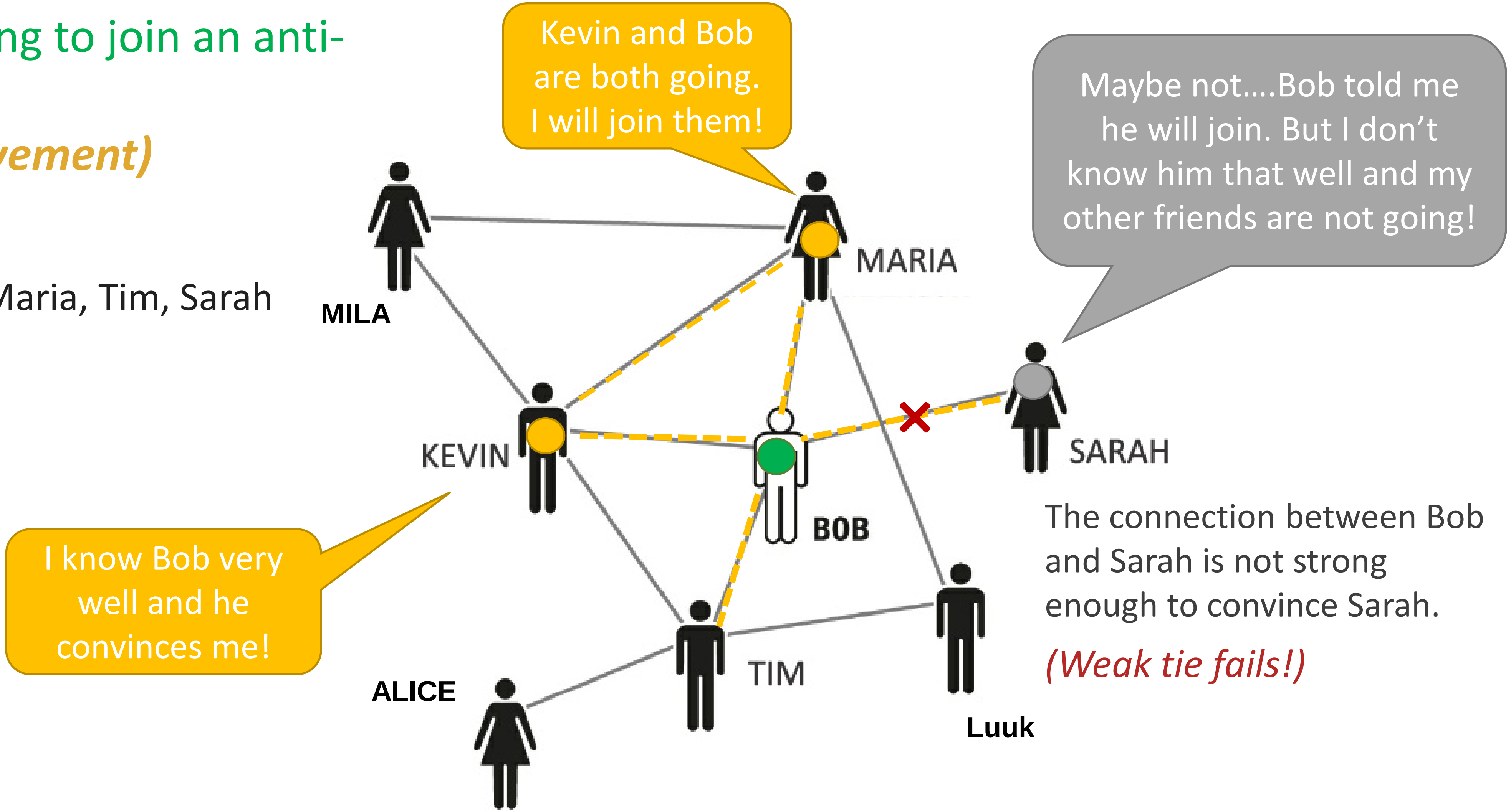


Demonstration of complex contagion in a small network

Bob is now considering to join an anti-lockdown protest
(High-risk social movement)

1st round: Bob

2nd round: Bob→ Kevin, Maria, Tim, Sarah



The connection between Bob and Sarah is not strong enough to convince Sarah.

(Weak tie fails!)

Thanks to the redundancy of strong ties, Kevin, Maria and Tim have multiple sources to confirm the credibility of the behavior and support each other.

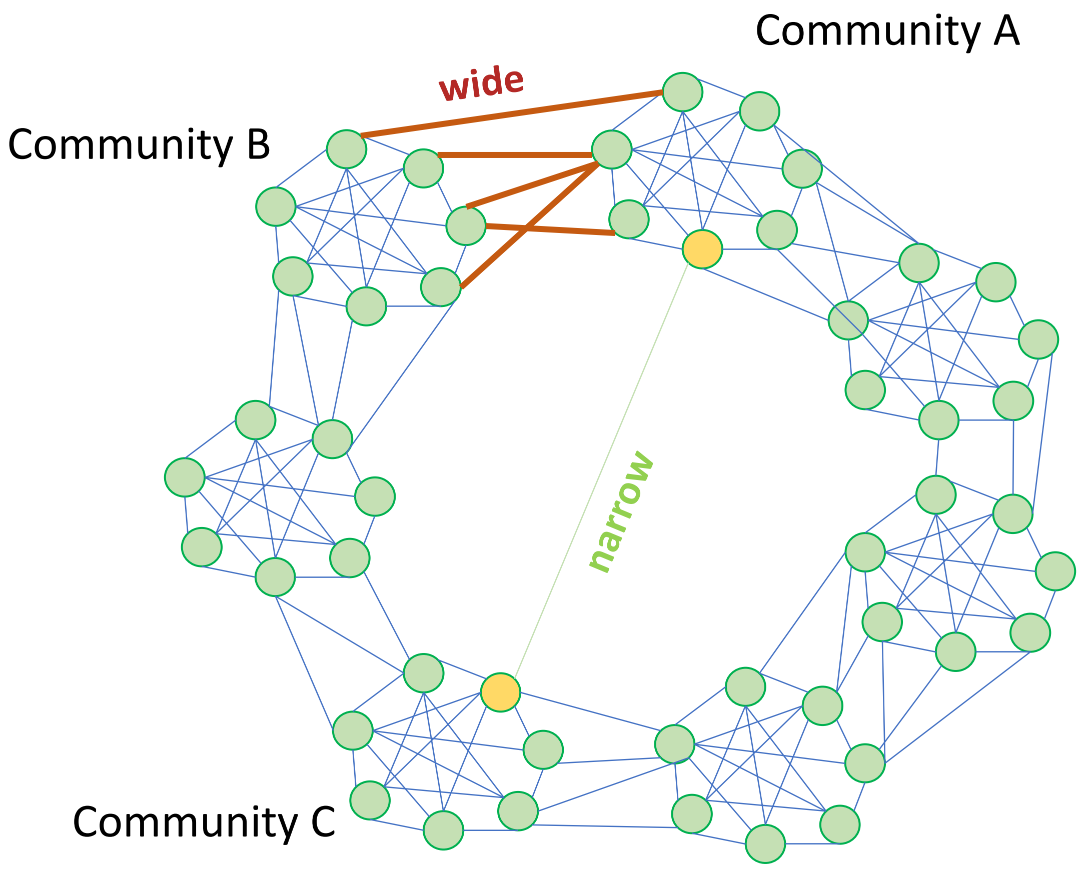
Complex Contagions and *the Weakness of Long Ties*, Centola &Macy, 2007

Complex contagions depend primarily on the width of the bridges across a network, not just their length.

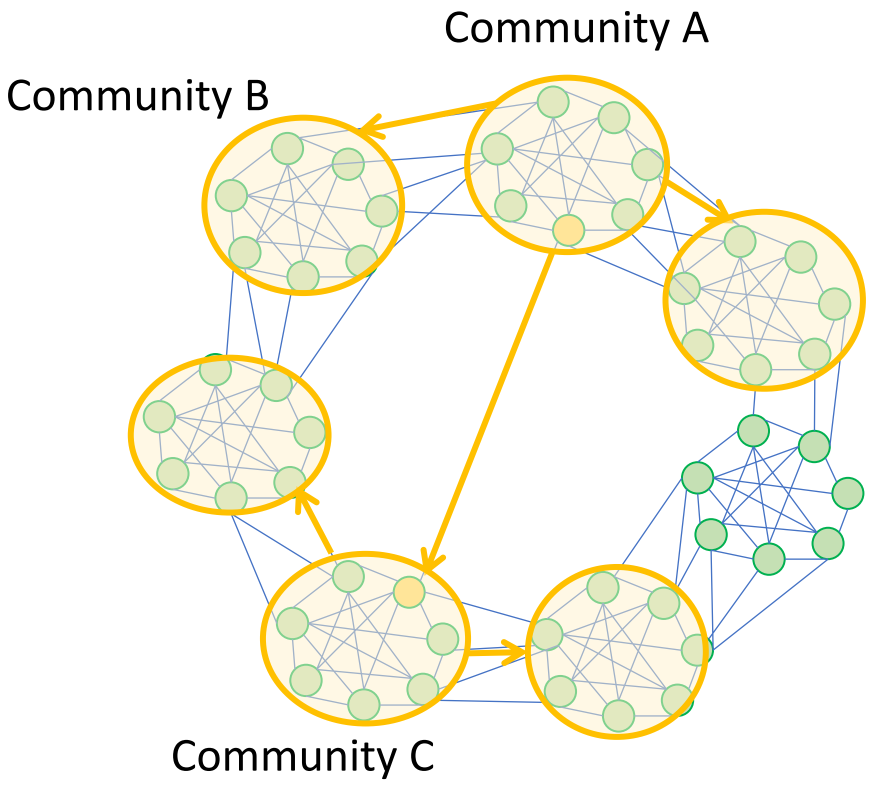
Damon Centola
https://www.damoncentola.com/

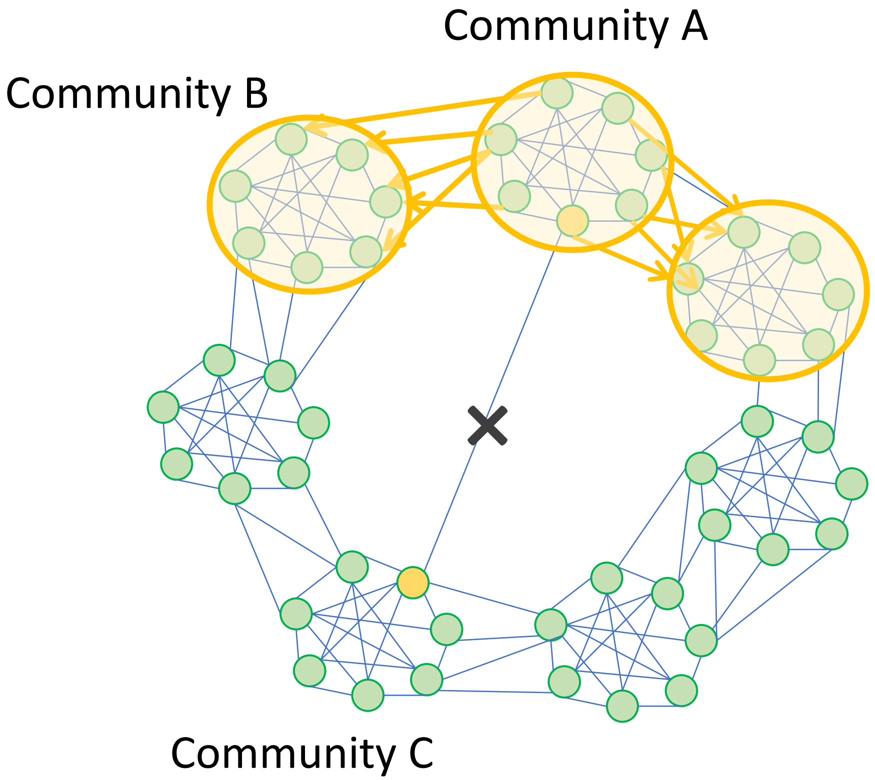
Width of a bridge decides the diffusion of complex behavior

- The width of a bridge between two communities is defined as the number of overlapping ties between them.



Diffusion of simple and complex contagion in the same network





Under simple contagion, the only tie between A and C *can* spread the rumor/virus to Community C and thus fasten the diffusion process.

Under complex contagion, the only tie between A and C *cannot* spread the new norm to Community C;

Complex contagion can only spread through multiple connections ("wide bridge") between nearby communities.

Today's programe

- **Simple contagion**

  Mechanism and the strength of 'weak ties'

- **Diffusion model for simple contagion**

  Independent cascade model and other variants

- **Complex contagion**

  Mechanism and the strength of 'strong ties'

- **Diffusion model for complex contagion**

  Threshold model and seeding

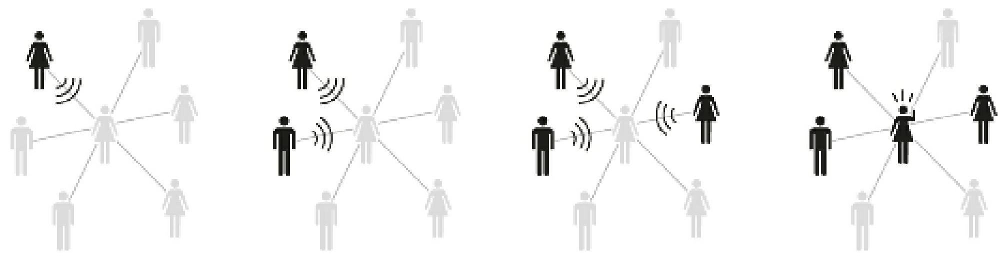
How to translate the complex contagion theory into numerical models?

Translate complex contagion into numerical models by “*Threshold*”

Threshold Models of Collective Behavior[1]

Mark Granovetter
State University of New York at Stony Brook

Threshold: To adopt a new behavior, an individual needs to be convinced by *an absolute number* or *a fraction* of his/her social contacts (e.g., the riskier the behavior the higher the threshold)



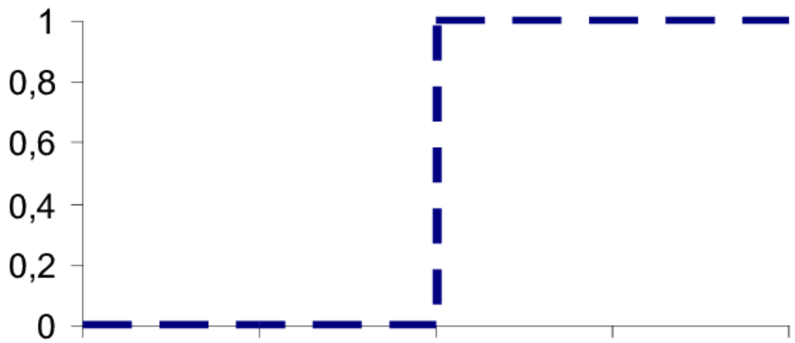
Formalization of a threshold model

- Node i can have two states — active ($S_i = 1$) and inactive (S_i =0); once activated, will remain active all the time.
- Imagine node i has **a fractional threshold** defined as θ_i, $0 \leq \theta_i \leq 1$
- Imagine node i has n direct neighbors.
- At each time step t, calculate the weighted status of node i's neighbors:

$$w_i(t) = \frac{Number\ of\ activated\ neighbors\ of\ node\ i\ at\ time\ t}{n}$$

$p_i(t)$ is the chance for node i to adopt the behavior at time step t:

$$p_i(t) = 1 |\ w_i(t) \geq \theta_i$$
$$p_i(t) = 0 |\ w_i(t) < \theta_i$$



- Stop when all the nodes are activated or the number of activated nodes is saturated.

Today's programe

- **Simple contagion**

  Mechanism and the strength of 'weak ties'

- **Diffusion model for simple contagion**

  Independent cascade model and other variants

- **Complex contagion**

  Mechanism and the strength of 'strong ties'

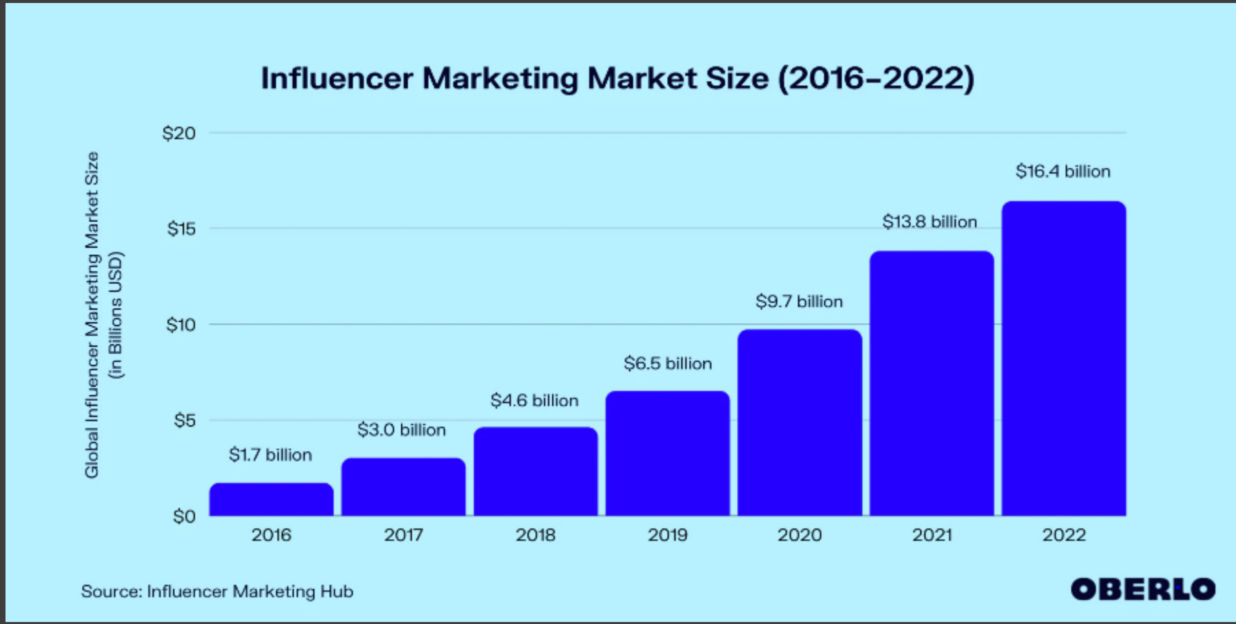
- **Diffusion model for complex contagion**

  Threshold model and seeding

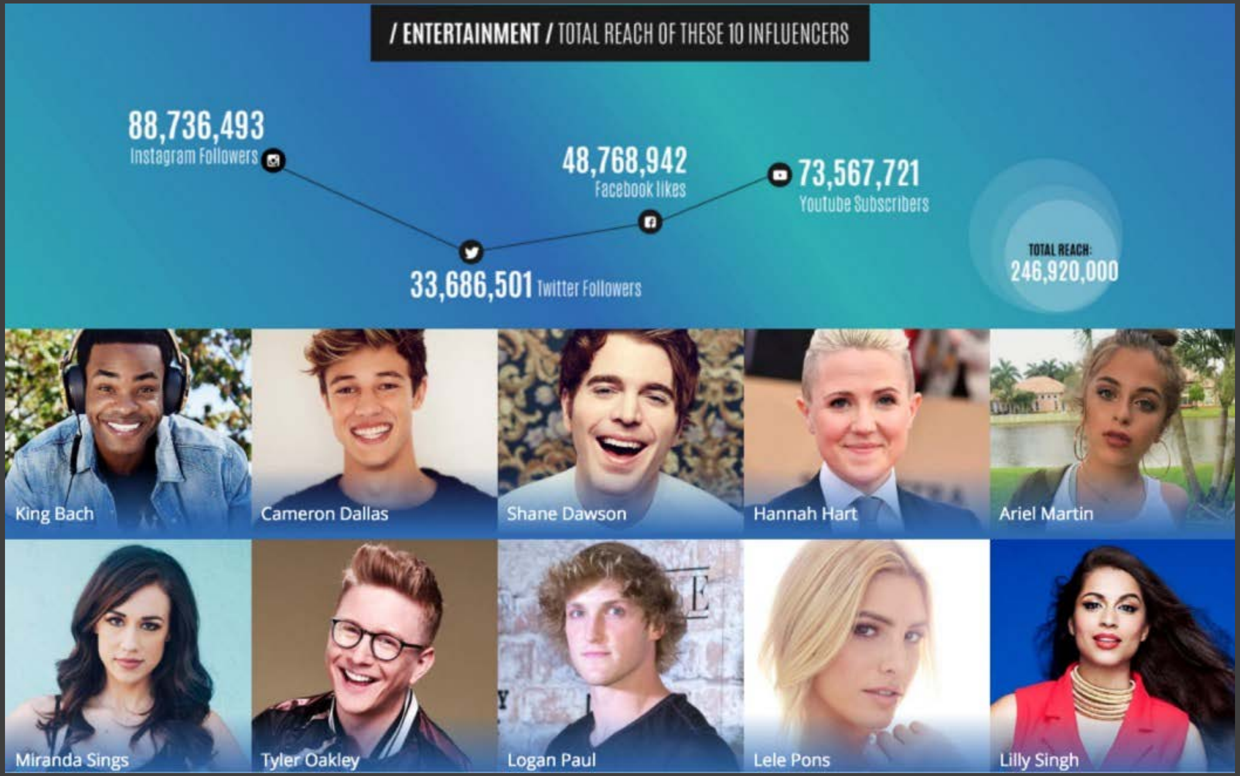
What kind of insights can you get from the model?

Influencer marketing industry: around $16.4bn in 2022

Almost half of marketers spent more than 20% of their budget on influencer posts



Are the ones with most followers always the best brand ambassadors?



Seeding: To find out a small set of influential nodes for activation/immunization

“Influencers”
“Super spreaders”
“Brand ambassadors”

Finding out this set of influential nodes is an influence maximization problem (IMP)

- Given a network $\boldsymbol{G(V, E)}$ with $\boldsymbol{V}$ and $\boldsymbol{E}$ respectively being the set of nodes and the set of links, there exists a function $\boldsymbol{f(S)}$ from a subset $\boldsymbol{S} \subseteq \boldsymbol{V}$ to a real number
- Influence maximization problem (IMP): To find a set $\boldsymbol{S}$ of a given size k (usually, $k \ll n = |\boldsymbol{V}|$) that maximizes/minimizes $\boldsymbol{f(S)}$.

Example of $\boldsymbol{f(S)}$:

With limited marketing budget, to find a small group of customers to offer discounts to eventually maximize total sales→ $f(S)$: the number of people who use the product

Obtaining an exact solution of IMP is very challenging

f(S): For a network of 60 people, to find out 5 people as the "seeds" to maximize the diffusion

- The number of possible combinations is $\binom{60}{5} = 27{,}307{,}560;$
- Run the threshold model for 27million times to find out which combination has the best outcome.

Using centrality heuristics to approximate the exact solution

Centrality heuristics:

- To rank all nodes according to their degrees or other centrality measures and pick up the k top-ranking nodes
- Naive but widely used due to simplicity (or used as references for other advanced algorithms)

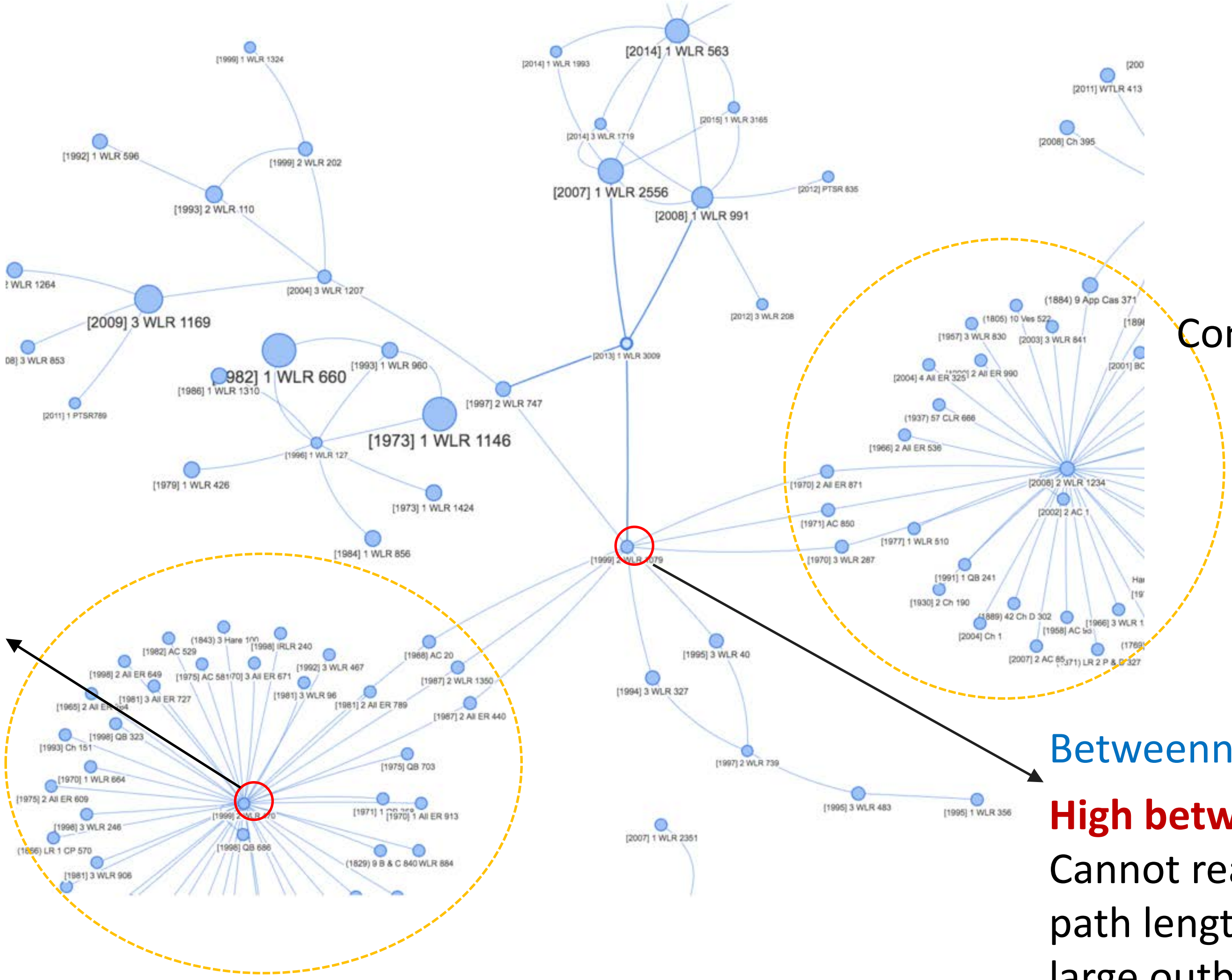
Top-k nodes from different measures:

Degree centrality; Local Rank (Neighbourhood-based)

Closeness centrality; Betweenness centrality (Path-based)

Eigenvector centrality (Position-based)

The choice of centrality heuristics depends on contagion mechanism



Community 2

Degree heuristic

High degree node ('hub');
Can reach many people within one path length

Community 1

Betweenness heuristic

High betweenness node;
Cannot reach many people within one path length, but can lead to a potentially large outbreak

Other sophisticated algorithms inspired by centrality heuristics

(Top-k nodes: sometimes inefficient since nodes of highest centrality may be highly clustered)

- High Degree Adaptive (HDA): First choose the node of the largest degree, then recalculate the degree of nodes after every step of node removal;
- Top-k spreaders from different communities: Network is divided into many communities by community detection algorithm. Then all communities are ranked in decreasing order according to their sizes. The first spreader is selected from the largest community according to a certain centrality index (e.g., to choose the node with the highest degree). Similarly, the node with the largest centrality index in the second largest community and having no edges incident to the previous communities.
- LocalRank: Consider the number of neighbors within path length 4

Using greedy algorithm to approximate the exact solution

For a network of 60 people, in order to find out 5 people as the "seeds" to maximize the diffusion, try $\binom{60}{5} = 27{,}307{,}560$ times

Only try
60+59+58+57+56=290 times

Add nodes one by one to the target set, ensuring that each addition brings the largest increase of ***influence*** to the previous set (maximize the incremental influence)

- k is the number of size of the seed set $\boldsymbol{S}$;
- Start with an empty set of $\boldsymbol{S}=\{\}$;
- At each time step, scan all nodes to find the one $\boldsymbol{v}$ that maximizes $\boldsymbol{f}(\ \boldsymbol{S} \cup \{v\}\ )$ and then updates $\boldsymbol{S}$ by including v
- After k time steps, one gets the target set $\boldsymbol{S}$ containing k influential nodes.

Reduce the computational time by considering the incremental spread of the k nodes individually rather than combined

Approximation guarantee of greedy algorithm

- Greedy algorithm approximates to the optimum $S*$ within a factor of around 60% or more in most cases: A seed set whose spread will be at least 60% of the spread of the optimal seed set.

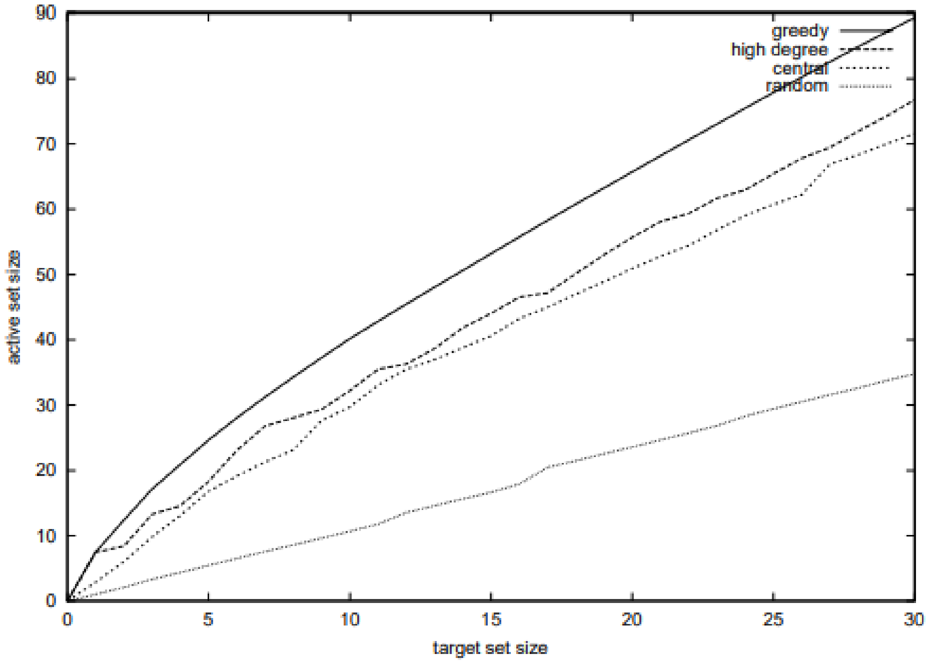


Figure 3: Independent cascade model with probability 1%

- Cost-effective lazy forward (CELF) algorithm: Improve greedy algorithm to further reduce the computational time.

D. Kempe, J. Kleinberg, ´ E. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’03, ACM Press, 2003, pp. 137–146.

Recap

Does social change spread in the network the same way as viruses and information?

No.

As a network scientist, what can you do?

Translate the complex contagion theory into numerical models by introducing "Threshold"

What insights can you generate from the model?

- Manipulate the diffusion by reconstructing the network (e.g., reduce weak ties by imposing travel ban)
- **Manipulate the diffusion by finding the right persons to activate/immunize: Suggestions on the influential subset**