

Network Analysis (INFOMNWA-2021)

Lecture 8: Influence manipulation

Jiamin Ou

What's the role of network science behind COVID-containing strategies?

Social-
distancing

Social bubbles

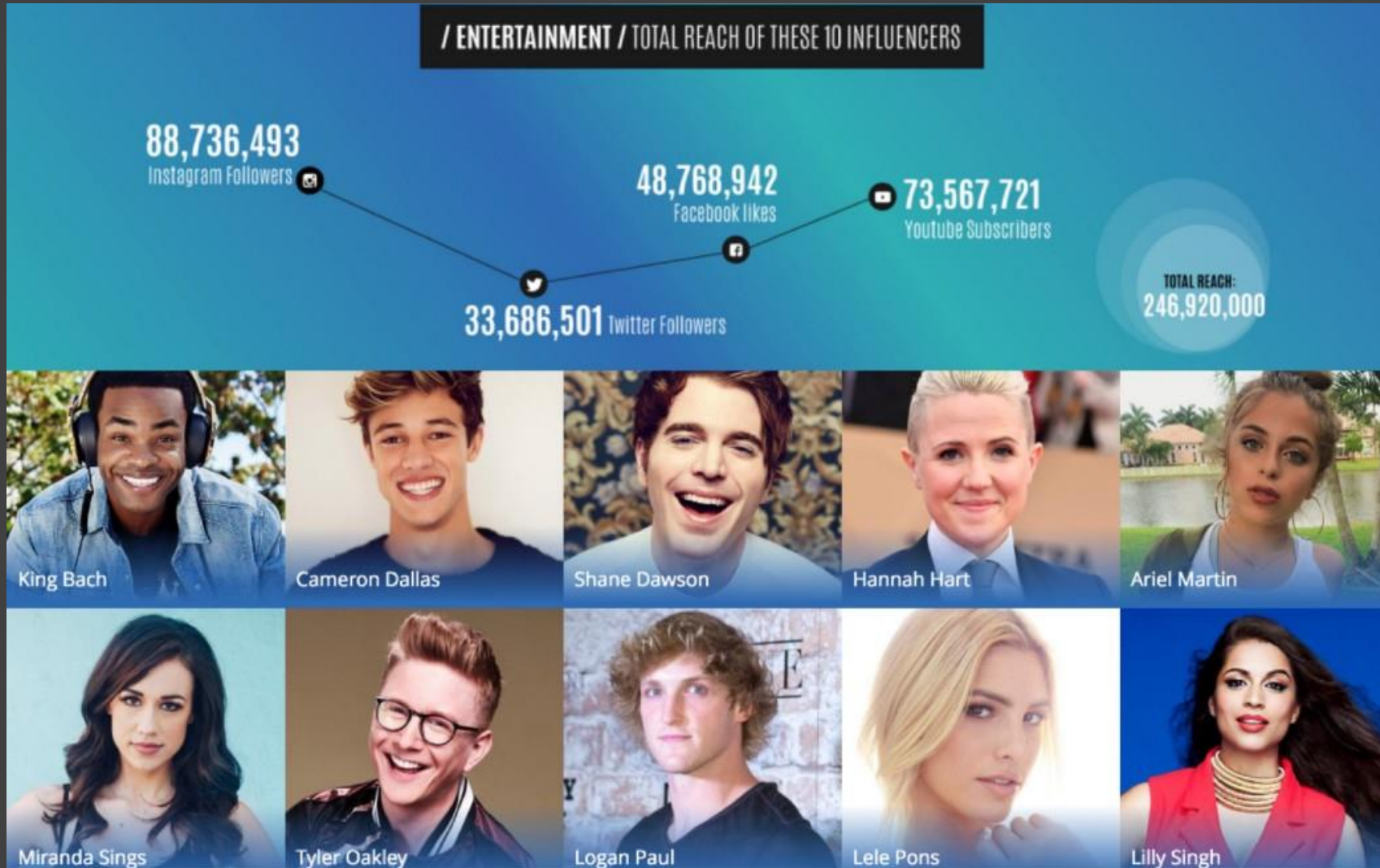
Super
spreaders

How did scientists come up with these measures and evaluate the effectiveness?



Influencer marketing industry: around \$6.5bn in 2019

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



Do the marketers spend the money wisely?

Are the ones with most followers the best ambassadors of your brand?

Today we try to answer

How can we model diffusion process in human network?

How can we find out a small set of influential nodes?



Influencers;
Super spreaders;
'Best ambassadors'

Today's program

- **Diffusion models**

- Independent cascade model

- SIR and its variants

- Threshold model

} Simple contagion with variations in states and probabilities

- **Influence maximization problem (IMP)**

- Structural IMP (network topology)

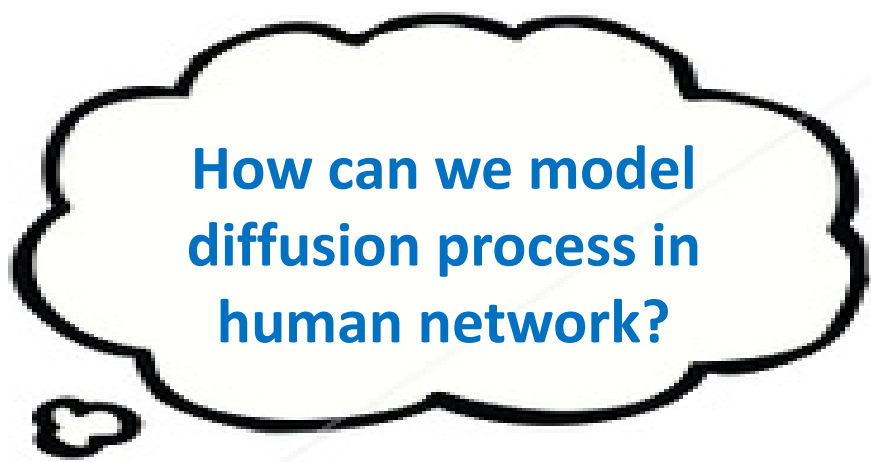
- Functional IMP (network topology + diffusion dynamics)

- **Approximation of IMP solutions**

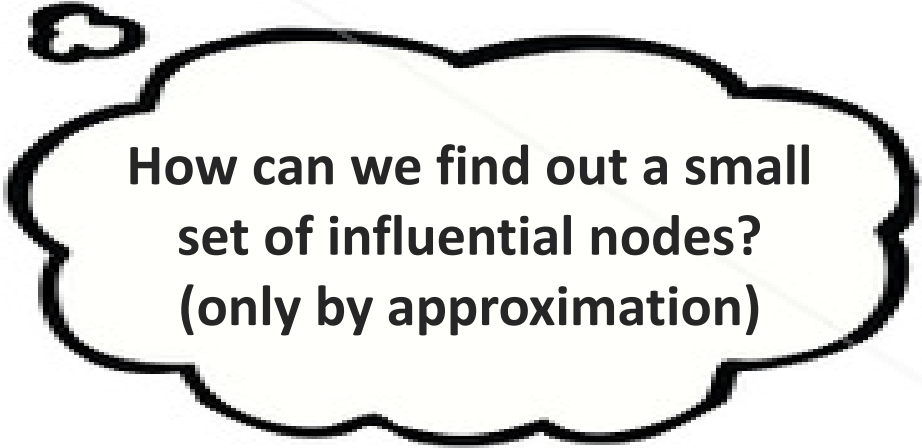
- Heuristics (degree, closeness, betweenness, k-shell, eigenvector....)

- Greedy algorithm

- CELF algorithm



How can we model diffusion process in human network?



How can we find out a small set of influential nodes? (only by approximation)

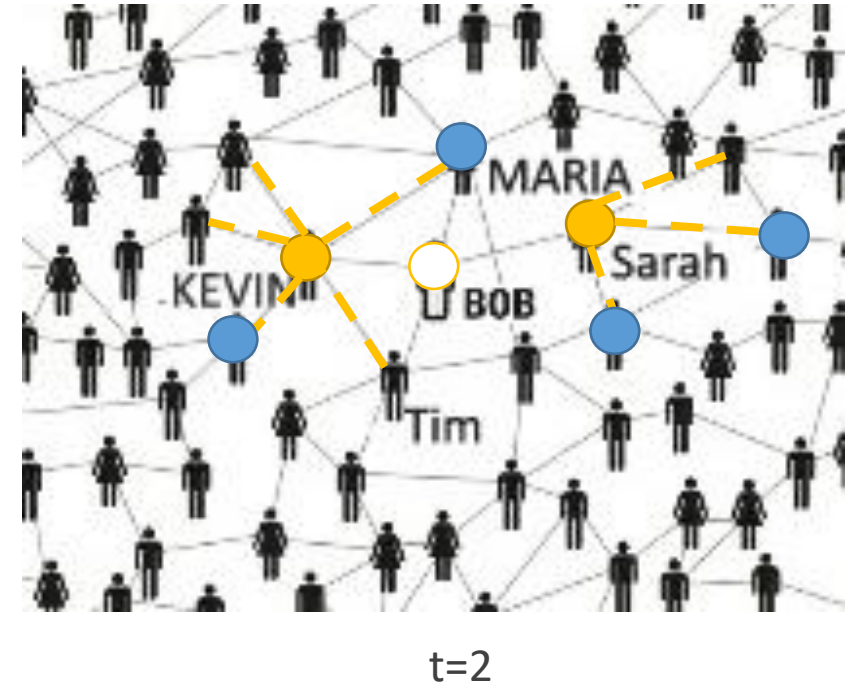
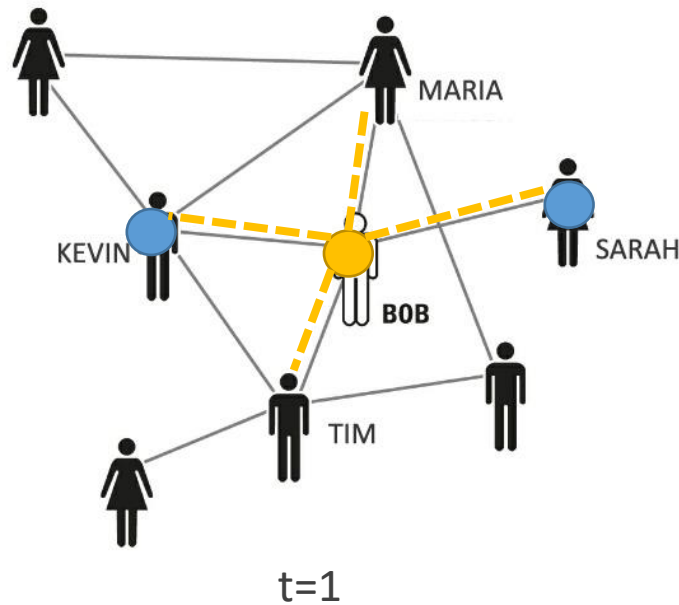
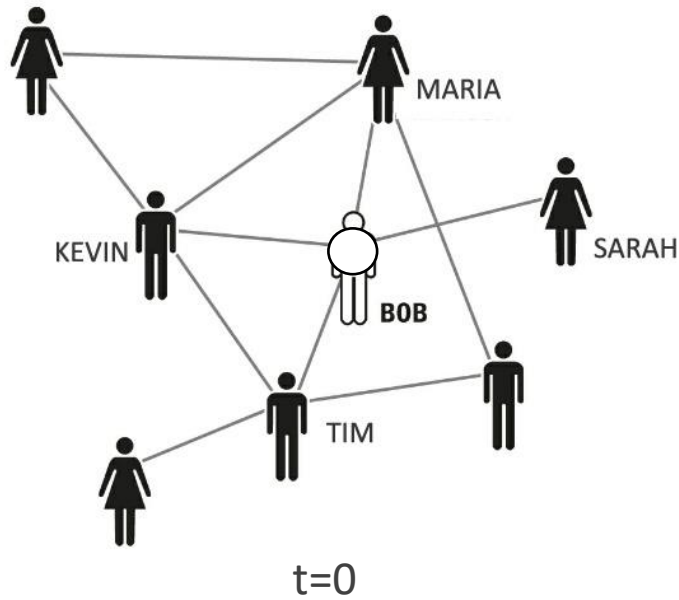
Now Bob tries out a new product and loves it.
He shares with other friends.



Independent Cascade Model

- Nodes can have two states — active ($S=1$) and inactive states ($S=0$); once activated, can not be inactive again (e.g., Bob used the product)
- At time $t=0$, k nodes are selected (i.e., activated). These nodes are called seed nodes
- When any node u is activated at time t , *it has a single chance* (Bob won't talk about the same product with his friends again and again) to activate each of its neighbors v at time $t+1$. The success depends on the probability p_{uv} assigned to the edge connecting u and v . (p_{uv} can be same for every edge or different by edges)
- Stop when all the nodes are activated or the number of activated nodes are saturated.

$p_{uw}=60\%$ for all edges



All nodes are inactive;
Bob is activated.

Check out the neighbours of Bob
 Bob-Kevin: Generate a random number, if **smaller** than 60%, Kevin activated;
 Bob-Tim: Generate a random number, if **larger** than 60%, Tim is not activated;
 Bob-Sarah: Generate a random number, if **smaller** than 60%, Sarah activated;
 Bob-Maria: Generate a random number, if **larger** than 60%, Maria is not activated;

Bob cannot activate his neighbors anymore;
 Check out the neighbors of Kevin and Sarah.

Susceptible-Infected- Recovered (SIR) Model

- People in a network (size= N) can have three states:

Susceptible (S): healthy people that can catch the virus with the contact of infected people, with certain probability β ;



Infected (I): people who have been infected and are capable of infecting susceptible individuals. (*being infectious within $1/\gamma$ time steps*)

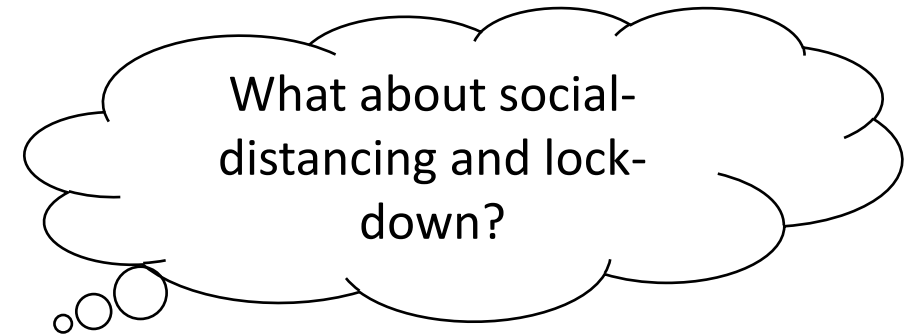


Recovered (R): people who have been infected and have either recovered from the disease and entered the removed compartment, or died.

- Virus “strength”: β/γ

Recovery rate
Not much you can do

Lower transmission rate
e.g., hand washing, mask wearing...



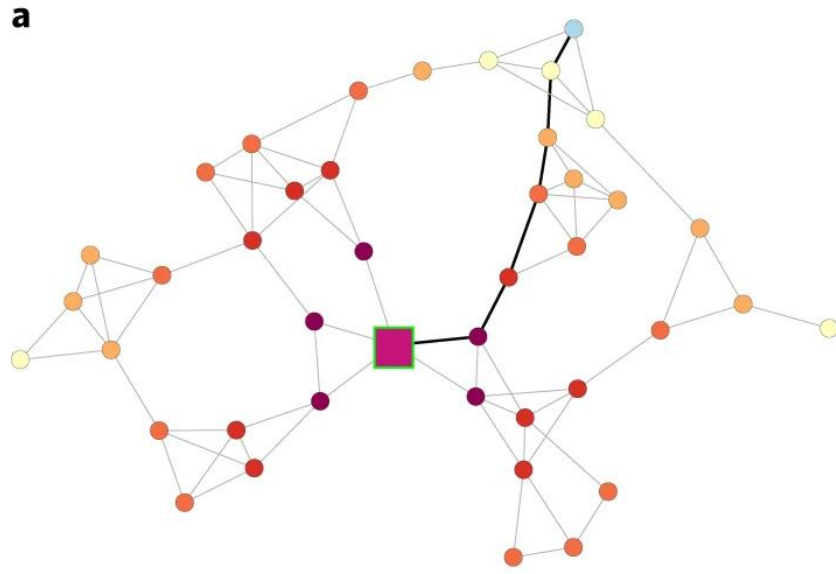
Diffusion dynamics + network topology = =?

Social network-based distancing strategies to flatten the COVID-19 curve

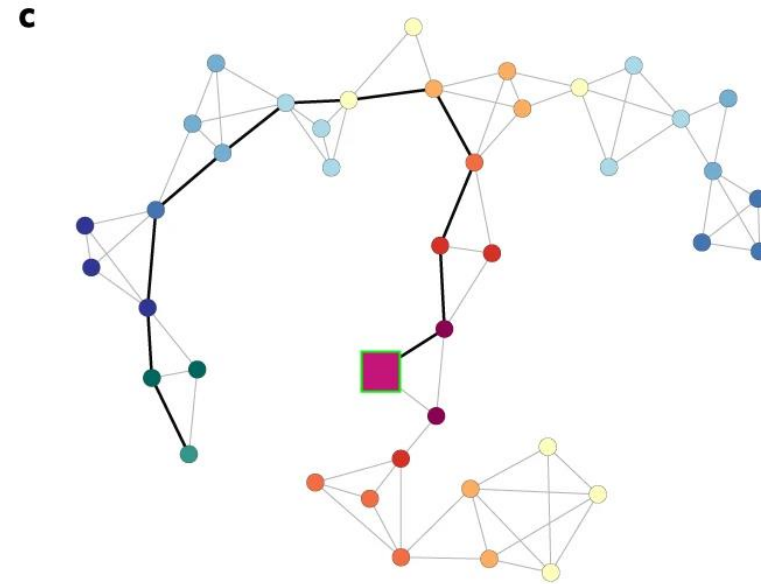
A diffusion model for COVID: Susceptible (S), expose (infected but not yet infectious) (E), infectious (I), recovered (R)

- Used the pre-COVID contact data to build a social network
- An infectious individual infects a healthy person when they interact, who then becomes exposed. This contagion occurs with probability $\pi_{infection}$.
- After a fixed number of steps $T_{exposure}$, an expose individual becomes infectious.
- After become infectious, recovery occurs with $T_{recovery}$ steps. Once recovered, individuals can no longer be infected.
- The process ends once there is no longer anyone exposed or infectious.

Start with two very small networks with same number of nodes and edges (i.e., human interaction intensity are the same)



(a) shorter average path lengths



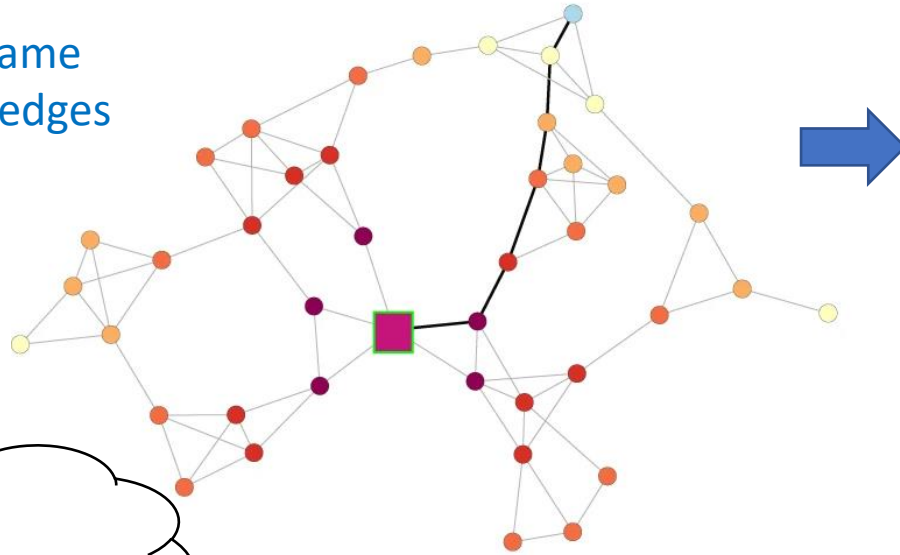
(c) longer average path lengths

Which network will see a faster spread of COVID?

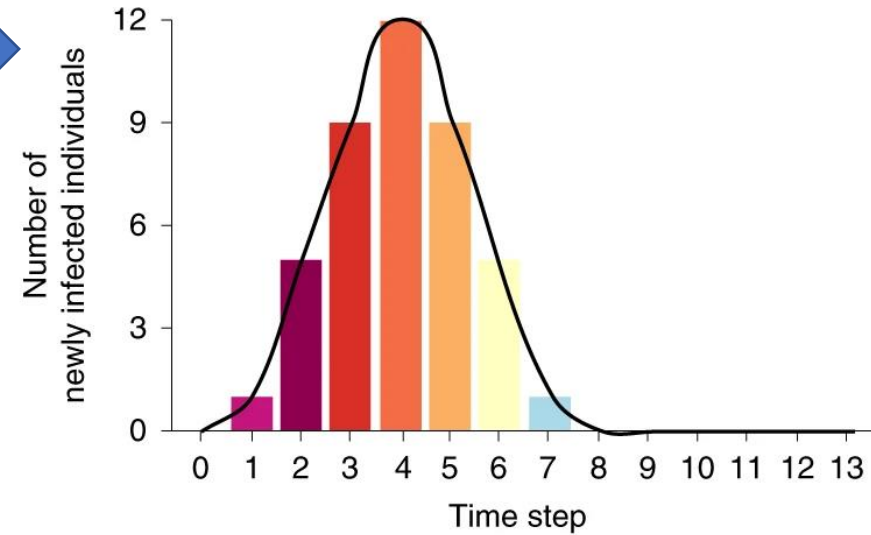
Network node colour indicates at which step a node is infected and maps onto the colours of the histogram bars

Two networks with same number of nodes and edges

How can this simple idea be used to construct social distancing strategies?

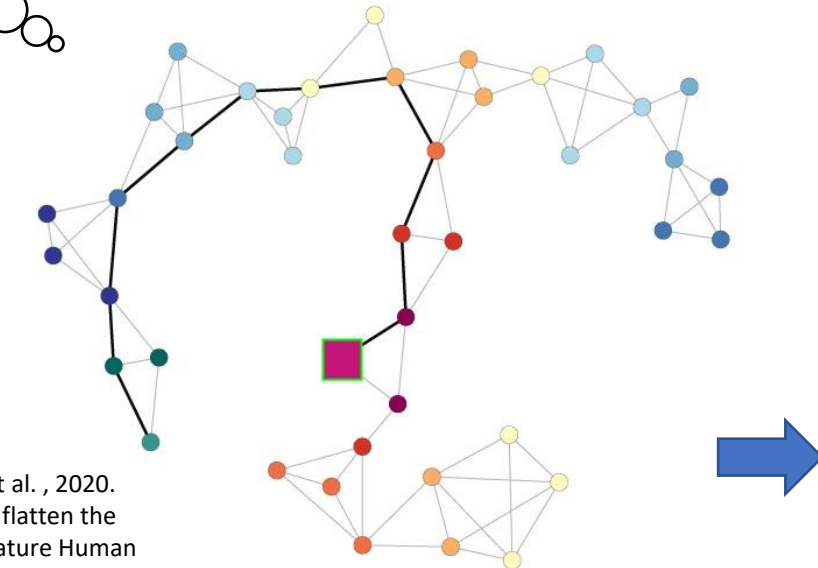
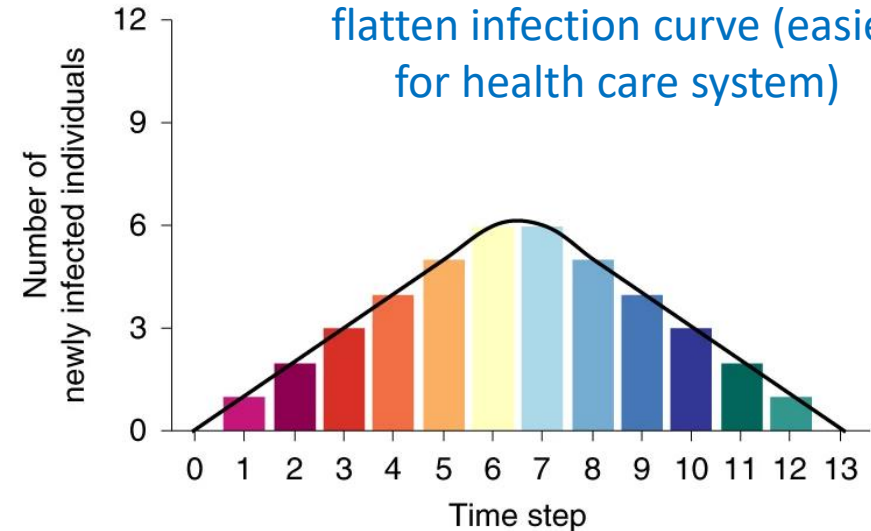


b

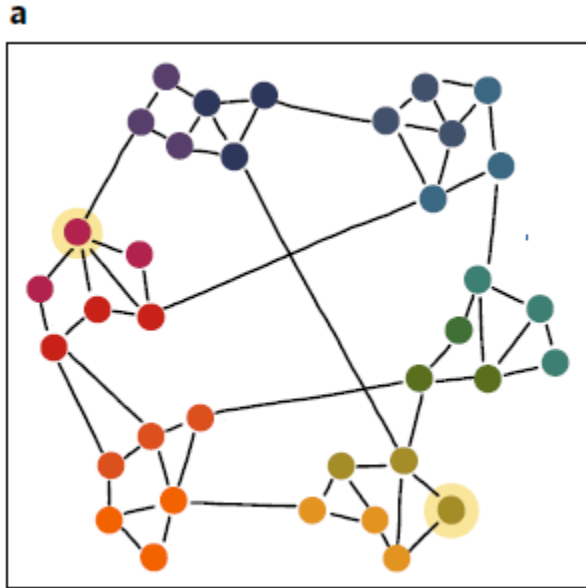


d

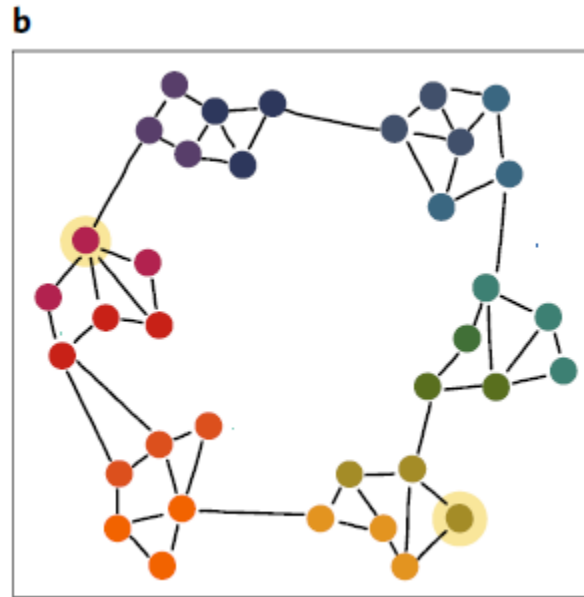
Takes longer and results in a flatten infection curve (easier for health care system)



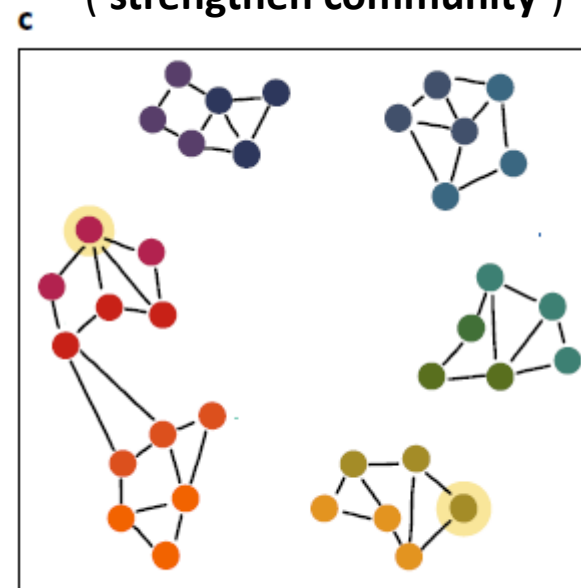
(a) **Pre-COVID** contact network with small-world properties



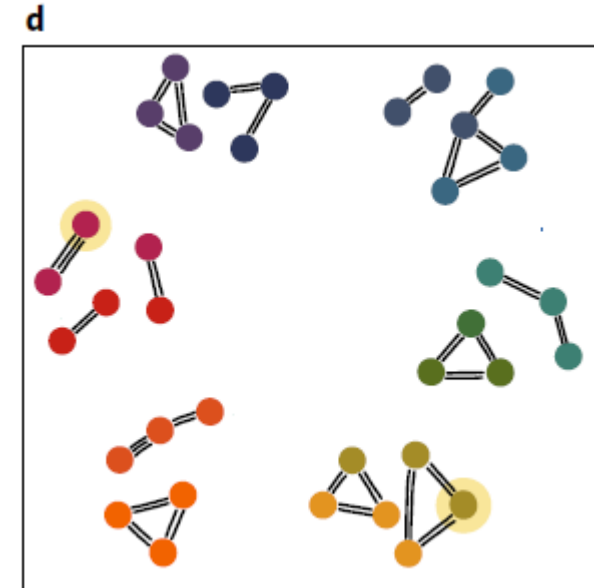
(b) Remove weak/long ties between distant communities ('seek similarity')



(c) Keep ties with redundancy but remove narrow bridges ('strengthen community')



(d) Build **bubbles** through repeated contact



Node colour represents an individual characteristic; Similar colour: people in the same neighbourhood (geographical) or those of similar income or age (socio-demographic)

Interact with those who are geographically proximate (e.g., living in the same neighbourhood) or those with similar characteristics (for example, age).

Two friends only meet when they have many friends in common (keeping contact in cohesive community)

Social network-based distancing strategies to flatten the COVID-19 curve

A diffusion model for COVID: Susceptible (S), expose (infected but not yet infectious) (E), infectious (I), recovered (R)

Reflect the change of the network typology under different distancing strategies

- Used the pre-COVID contact data to build a social network
- At each step, one individual is picked at random and initiates an interaction with the probability π_{contact} . ($\pi=1$ for pre-COVID; 0.5 for 'seek similarity', 'strengthen community' and 'social bubble')
- An actor initiating an interaction can only pick one interaction partner, under some preferences (reflecting different social distancing strategies).
- An infectious individual infects a healthy person when they interact, who then becomes exposed. This contagion occurs with probability $\pi_{\text{infection}}$.
- After a fixed number of steps T_{exposure} , an expose individual becomes infectious.
- After become infectious, recovery occurs with T_{recovery} steps. Once recovered, individuals can no longer be infected.
- The process ends once there is no longer anyone exposed or infectious.

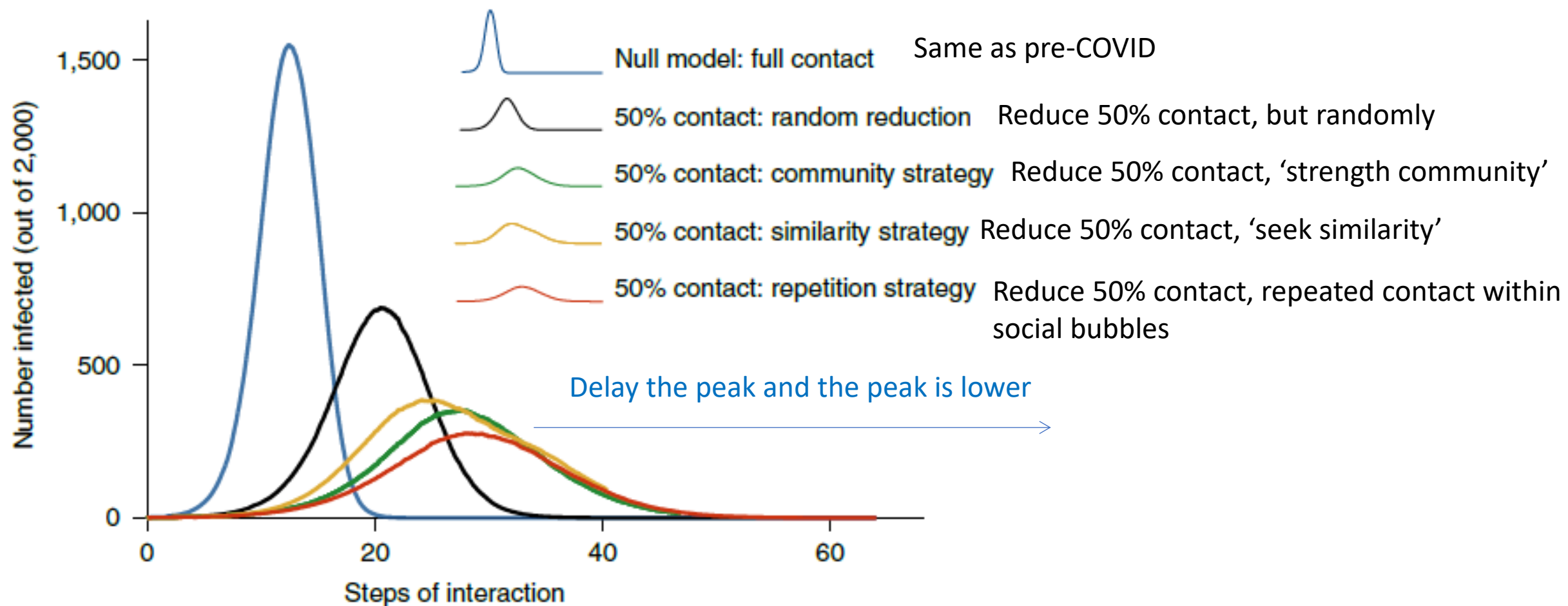


Fig. 4 | Average infection curves. Curves compare four contact reduction strategies with the null model of no social distancing. The underlying network structure includes 2,000 actors and the benchmark network characteristics described in the main text.

Today's program

- **Diffusion models**

 - Independent cascade model

 - SIR and other variants

 - Threshold model (complex contagion)

- **Influence maximization problem (IMP)**

 - Structural IMP (network topology)

 - Functional IMP (network topology + diffusion dynamics)

- **Approximation of IMP solutions**

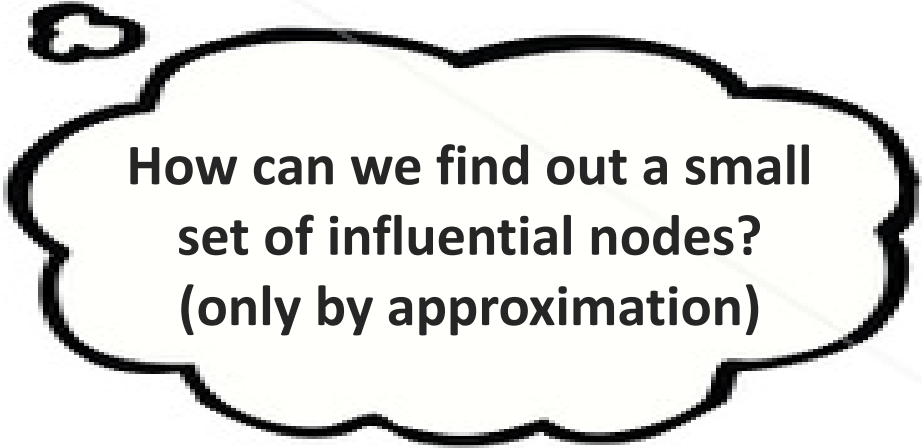
 - Heuristics (degree, closeness, betweenness, k-shell, eigenvector....)

 - Greedy algorithm

 - CELF algorithm



How can we model
diffusion process in
human network?



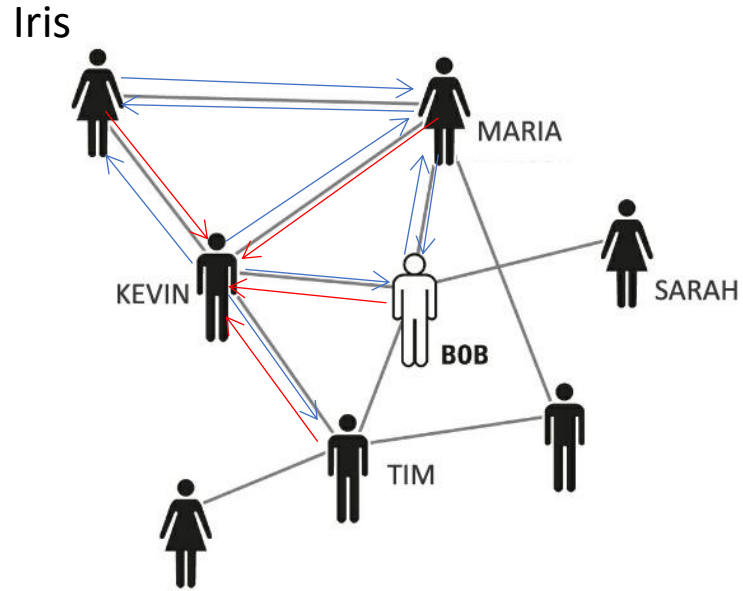
How can we find out a small
set of influential nodes?
(only by approximation)

Linear threshold model

- Nodes can have two states — active ($S=1$) and inactive states ($S=0$); once activated, will remain active all the time.
- Influence comes only from one's direct neighbors $N(i)$ nodes, w_{ji} influence $j \rightarrow i$
- Require $\sum_{j \in N(i)} w_{ji} \leq 1$
- Each node has a acceptance threshold from $\theta_i \in [0,1]$
- In each time step, each inactivated node reviews the status of all his/her direct neighbors $N(i)$ and will be activated if the weighted fraction of active nodes exceeds threshold

$$\sum_{\text{active } j \in N(i)} w_{ji} > \theta_i$$

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



If Bob is activated at t_1 :
 $0.5 < 0.7$; Kevin remains inactive

If Tim is also activated at t_2 :
 $0.5 + 0.1 < 0.7$; Kevin remains inactive

If Maria is also activated at t_3 :
 $0.5 + 0.1 + 0.3 > 0.7$; Kevin is activated.

Let's focus on Kevin:

Influence from Bob to Kevin (w_{bk}): 0.5

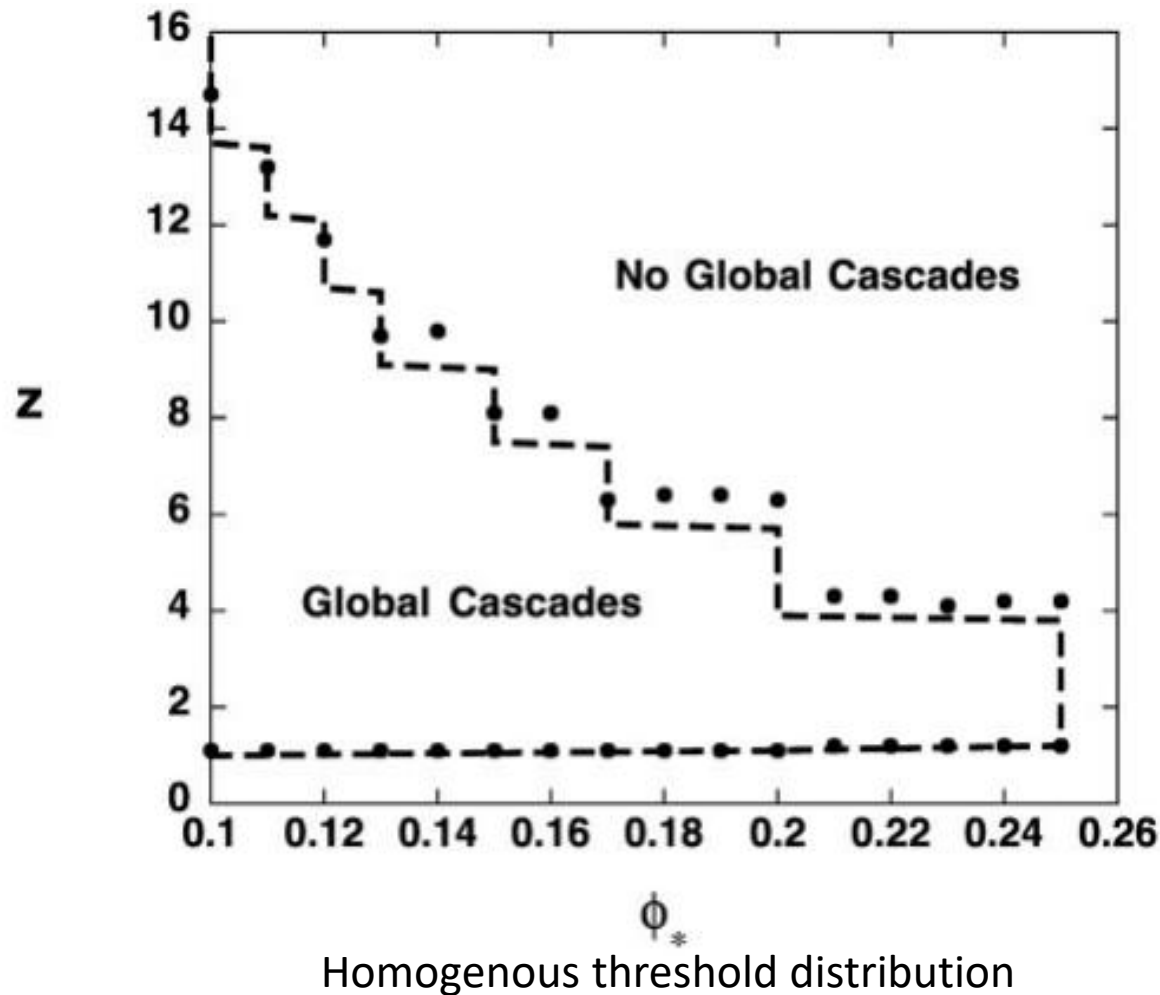
Influence from Tim to Kevin (w_{tk}): 0.1

Influence from Maria to Kevin (w_{mk}): 0.3

Influence from Iris to Kevin (w_{ik}): 0.1

And the threshold of Kevin $\theta_k = 0.7$

Global cascade (sufficiently large)
Triggered by a small set of nodes
Random graph ER
Z: average degree of the graph



Threshold values, threshold distribution, average degree and degree distribution \rightarrow successful cascade or not?

What might happen if threshold is different from nodes? (e.g., same mean degree but with hubs?)

Today's program

- **Diffusion models**

- Independent cascade model

- SIR and other variants

- Threshold model

- **Influence maximization problem (IMP)**

- Structural IMP (network topology)

- Functional IMP (network topology + diffusion dynamics)

- **Approximation of IMP solutions**

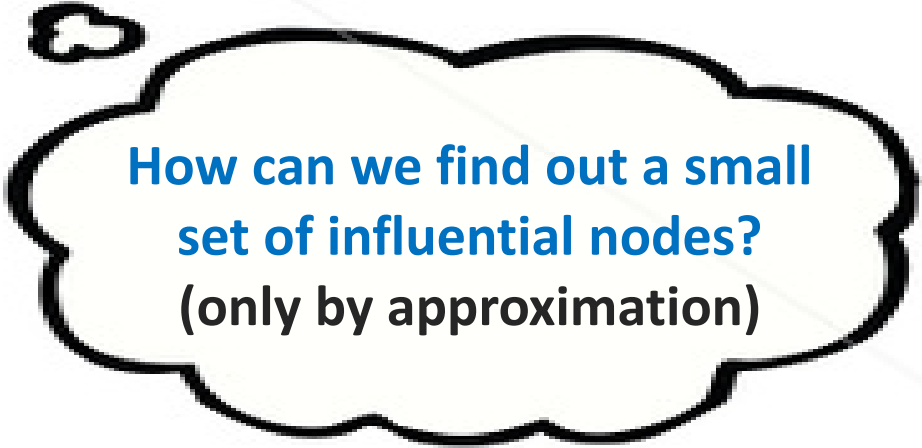
- Heuristics (degree, closeness, betweenness, k-shell, eigenvector....)

- Greedy algorithm

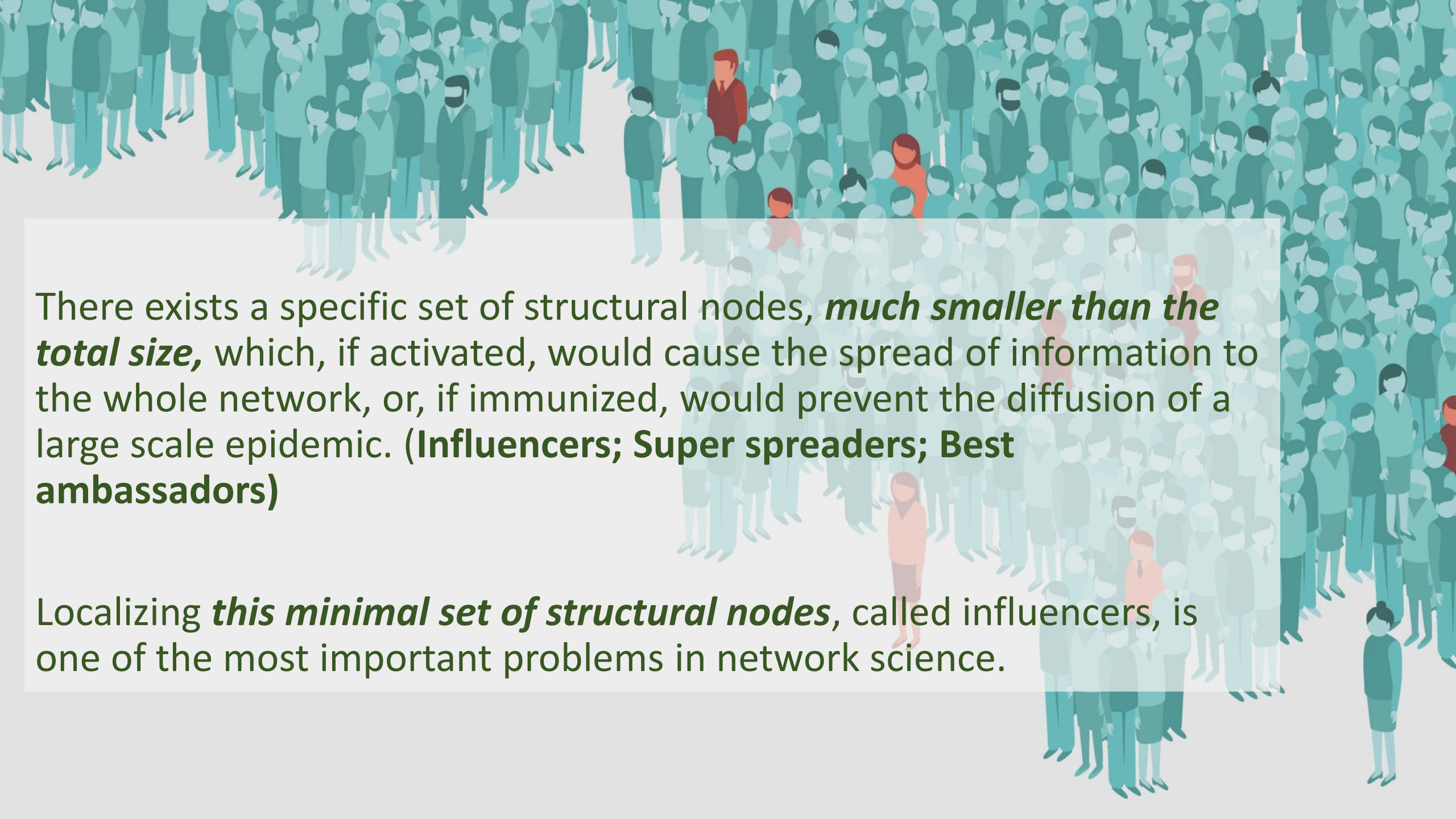
- CELF algorithm



How can we model
diffusion process in
human network?



How can we find out a small
set of influential nodes?
(only by approximation)

The background of the slide is a dense crowd of stylized human figures. Most figures are teal, while a few are red, representing a network of people. The figures are arranged in a way that suggests a large gathering or a complex network structure.

There exists a specific set of structural nodes, ***much smaller than the total size***, which, if activated, would cause the spread of information to the whole network, or, if immunized, would prevent the diffusion of a large scale epidemic. (**Influencers; Super spreaders; Best ambassadors**)

Localizing ***this minimal set of structural nodes***, called influencers, is one of the most important problems in network science.

Influence maximization problem (IMP)

- Given a network $G(V, E)$ with V and E respectively being the set of nodes and the set of links, if there exists a function $f(S)$ from a subset $S \subseteq V$ to a real number
- Influence maximization problem (IMP): to find the subset S with a given size k (usually, $k \ll n = |V|$) that maximizes $f(S)$.
- **Structural IMP:** the influence function $f(S)$ is fully determined by the network topology
- **Functional IMP:** the influence function $f(S)$ involves network topology and mechanisms or dynamical processes with a number of parameters that are independent to the topology

Structural IMP

The influence function $f(S)$ is fully determined by the **network topology**

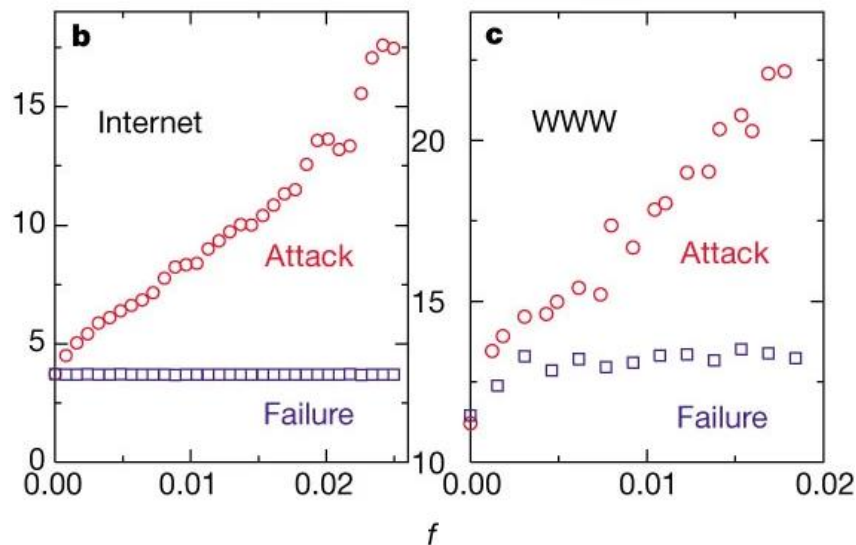
When you want to study the connectivity/robustness of a network

Examples:

$f(S)$: the diameter/average path length after the removal of S (a subset of nodes)

$f(S)$: the size of the largest component after the removal of S (a subset of nodes)

The diameter after removing 0~5% of random nodes (Failure) and highly connected nodes (Attack)



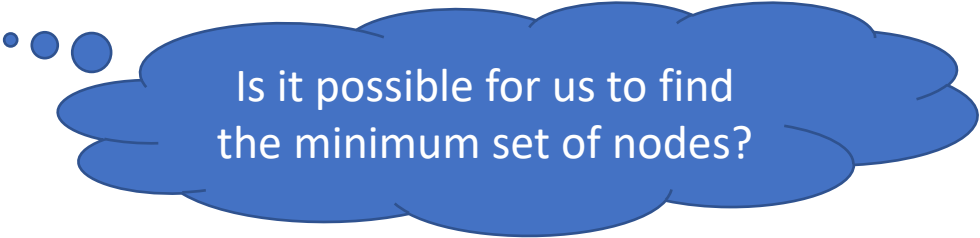
Functional IMP

$f(S)$ involves **network topology** and **mechanisms or dynamical processes** with a number of parameters that are independent to the topology

Examples:

Independent cascade model: With limited marketing budget, to find a small group of customers to offer discounts to eventually maximize total sales

SIR model: When confronting communicable diseases, to find the minimum set of immunized people, that would protect the whole population from being infected or slow the transmission to a certain rate



Is it possible for us to find the minimum set of nodes?

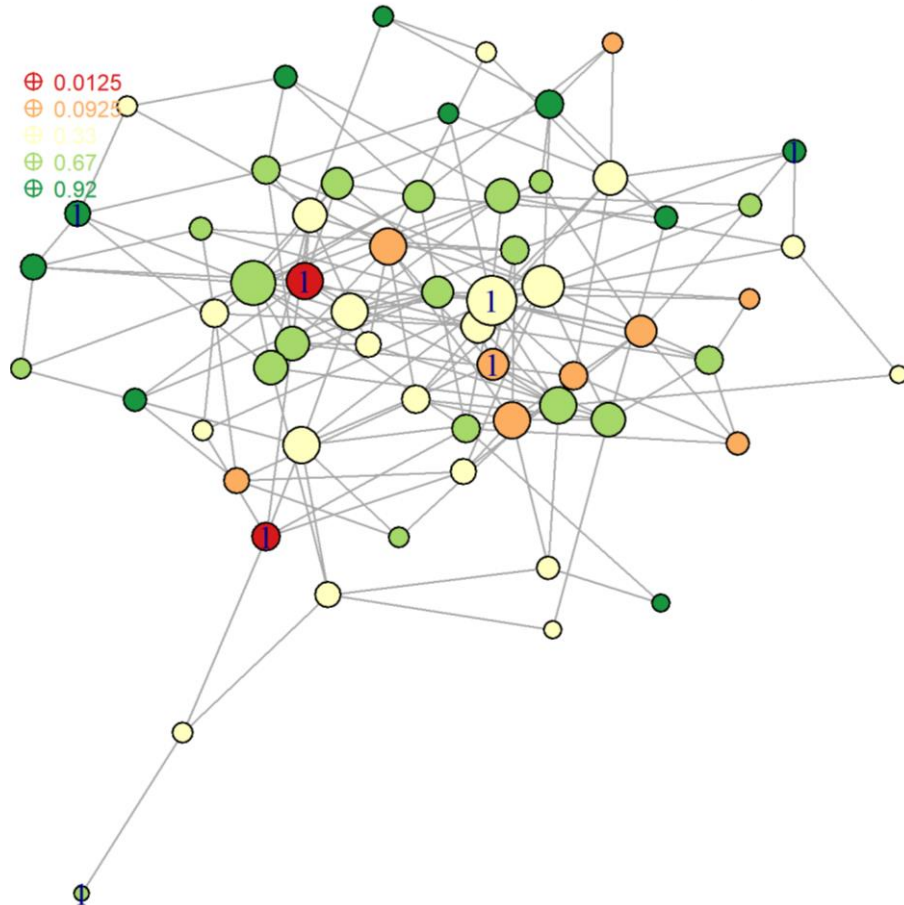
Dutch social housing legislation : Any significant community-wide alterations of housing units can only be made if more than 70% of the tenants approve. In some cases, this means that large groups of households need to be motivated before the usage of gas can be discontinued.

To find out the smallest k to activate at least 40 people in a 60 people building; using Linear threshold model

Exhaustively search all subsets: $\binom{60}{5} = 27,307,560$;

If 5 nodes cannot achieve the target of 40 people, need to search $\binom{60}{6} = 50,063,860$

.....



To find a minimal set of nodes that optimize a global function of influence (subset sum problem) is shown to be a NP-hard problem.

Today's program

- **Diffusion models**

 - Independent cascade model

 - SIR and other variants

 - Threshold model

- **Influence maximization problem (IMP)**

 - Structural IMP (network topology)

 - Functional IMP (network topology + diffusion dynamics)

- **Approximation of IMP solutions**

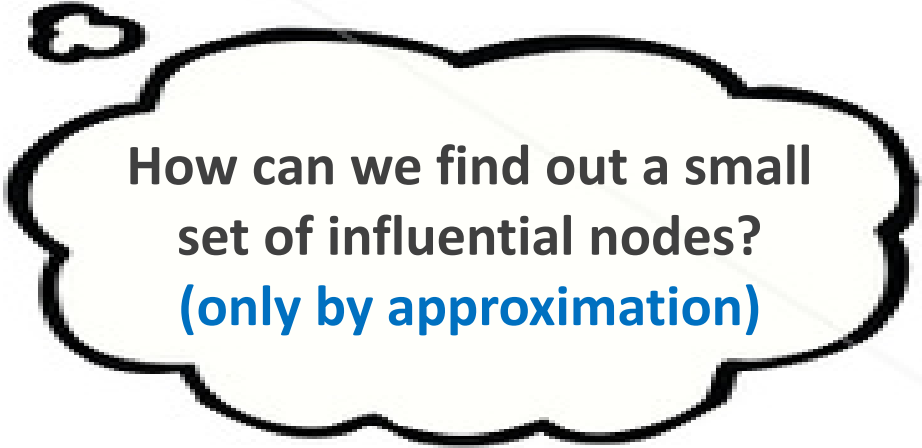
 - Heuristics (degree, closeness, betweenness, k-shell, eigenvector....)

 - Greedy algorithm

 - CELF algorithm



How can we model contagion in human network?



How can we find out a small set of influential nodes?
(only by approximation)

Approximation of IMP solution

Influence maximization problem: exact solution is hard to find

To find approximate solutions instead of the exact solution:

Degree centrality; Local Rank
(Neighbourhood-based)

Closeness centrality;
Betweenness centrality

(Path-based)

k-shell; Eigenvector centrality
(Position-based)

Top-k
nodes

Heuristic algorithms: to rank all nodes according to their degree or another centrality measure and directly pick up the k top-ranking nodes

Naive but widely used due to simplicity (or used as references for other more advance algorithms)

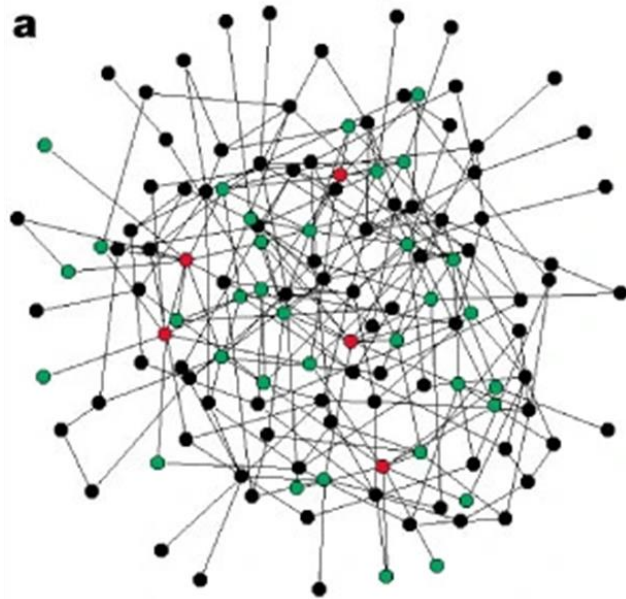
Degree centrality

The number of people that you reach within one step (“Neighborhood-based”)

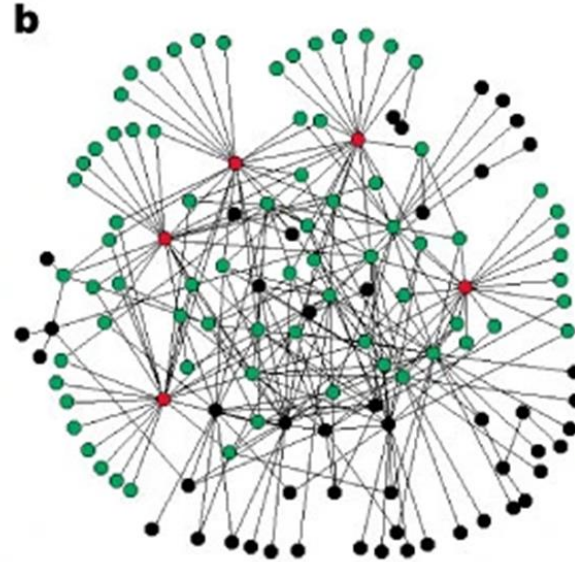
Red, the five nodes with the highest number of links

Green, their first neighbours

27% of the nodes are reached by the five most connected nodes



Random network by ER model



Scale-free network by BA model

> 60% are reached by the five most connected nodes

Both networks contain 130 nodes and 215 links

Degree centrality

- Focus on immediate effect (e.g., very limited time to see the output)
- Very easy to identify the highest degree nodes for both online and offline networks (The least demanding on network information)

The one with most followers in social media (online marketing)
The most popular one in a neighborhood (energy transition)
The occupations that meet most people (communicable diseases)

- Not ideal to maximize the effect of the whole network
- Effectiveness varies a lot (e.g., random network VS SF network)



Closeness and betweenness centrality

Based on the paths in a network:

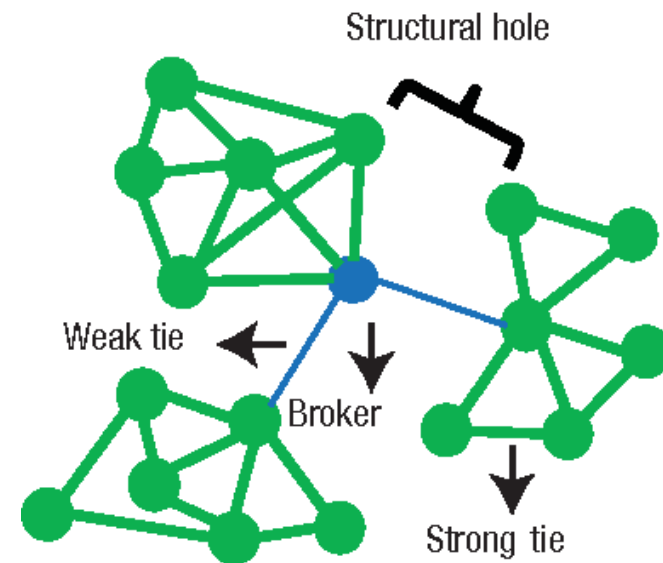
Closeness of node i : sum of the shortest path between node i and other nodes

Betweenness of node i : the number of times that node i lies along the shortest path between two other nodes

Closeness centrality: Reflect how efficient a node exchange information with all other nodes

(a good propagator, if you aim to target all people in the network)

Betweenness centrality: Reflect a node's potential power in controlling the information flow in a network ('broker' or the ones with lots of weak ties)



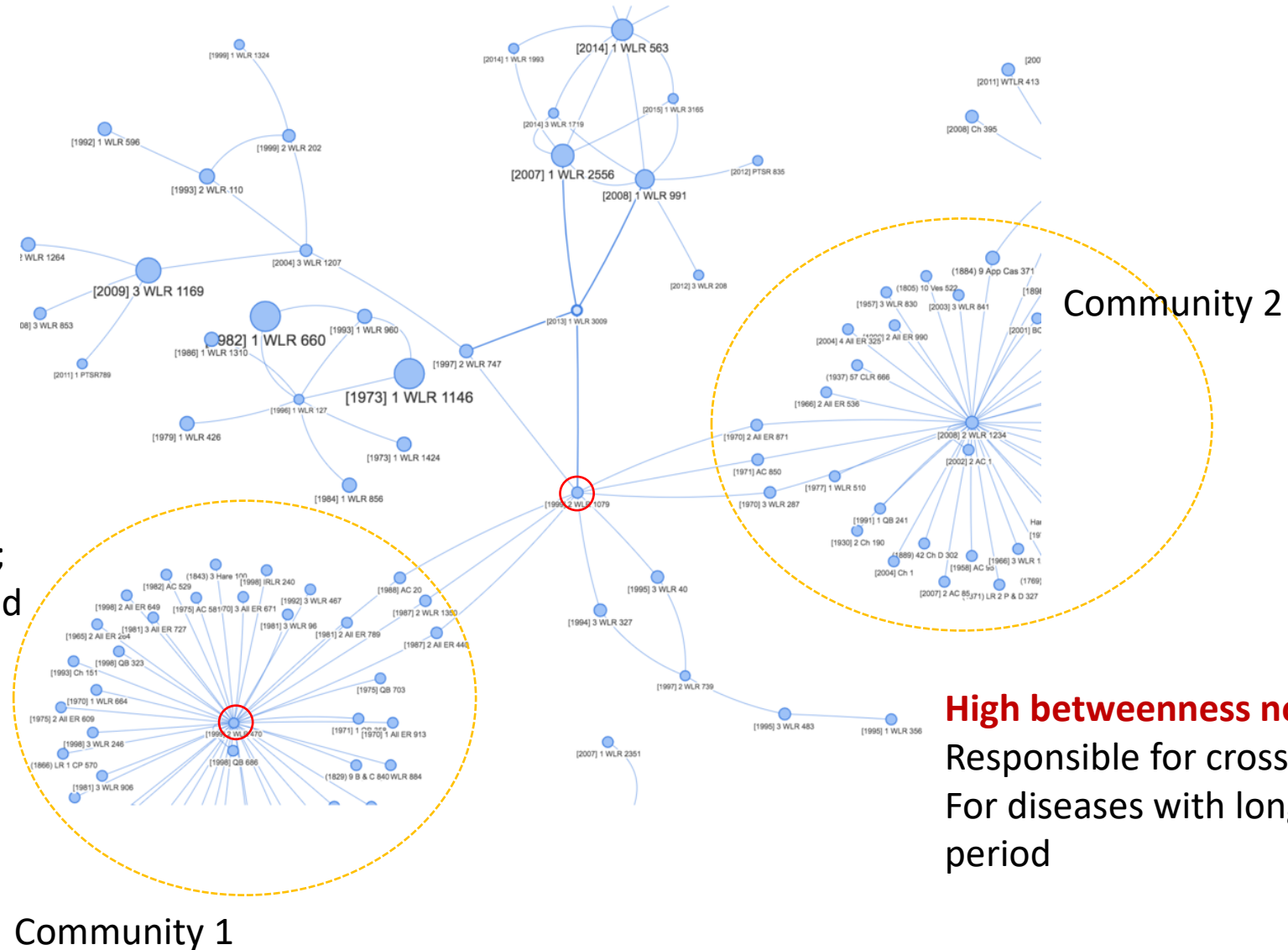
Closeness and betweenness centrality

- Reflect the structure of the whole network
- For project aiming for the whole population and believe that signal will not decay significantly even for the longest path

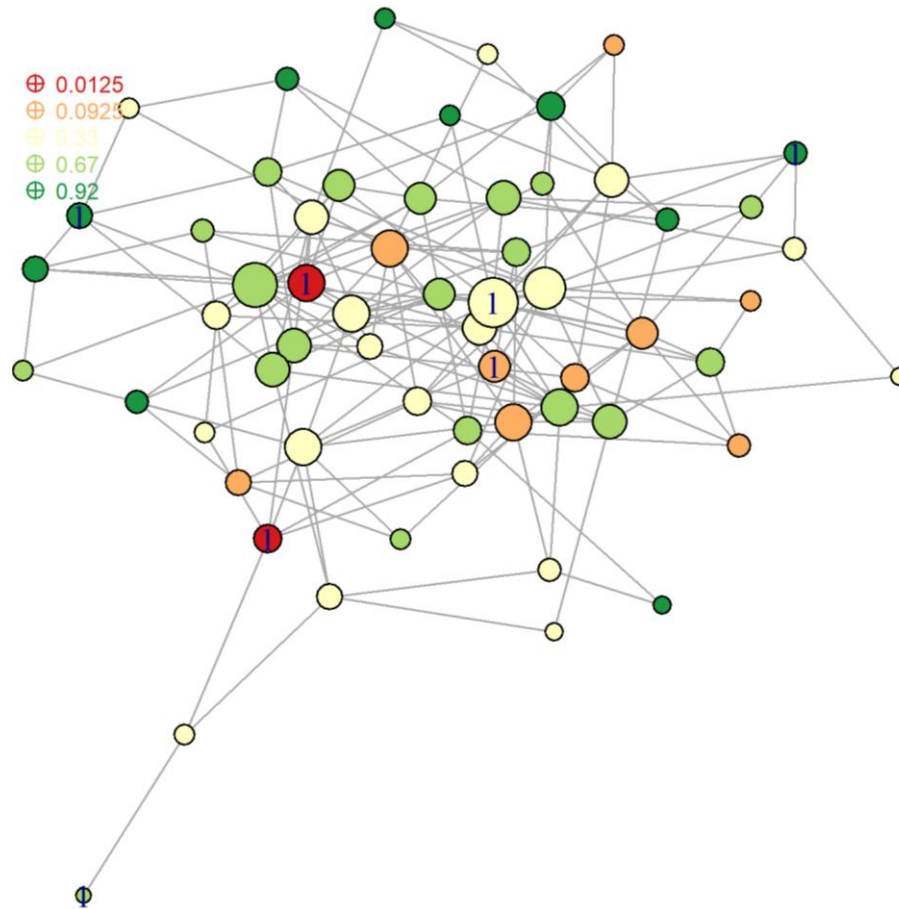
- Require complete information of the network
- Computationally expensive
- Not easy to interpret and extend the results from sample network (e.g., what are the demographic characteristics of people with high closeness in real world?)



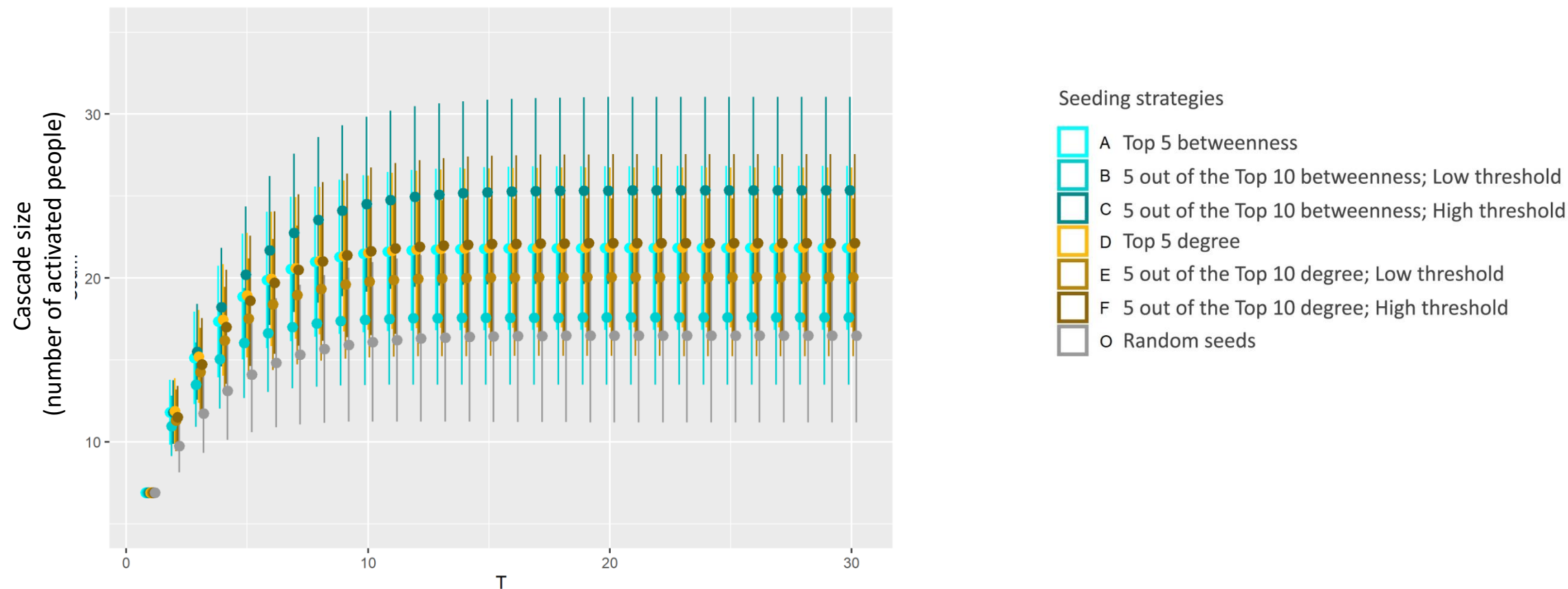
The highest degree and highest betweenness nodes in a contact network



Comparison of degree and betweenness heuristics in the small network of a social building



If we seed 5 people



Degree and betweenness heuristics perform better than random seeding;
but it is hard to tell which one is better.

k-shell and eigenvector centrality

Based on the position (not only the number of direct neighbors, but the importance of these direct neighbors) in a network.

- Reflect the structure of the whole network
- For project aiming for the maximum increase within a few time steps



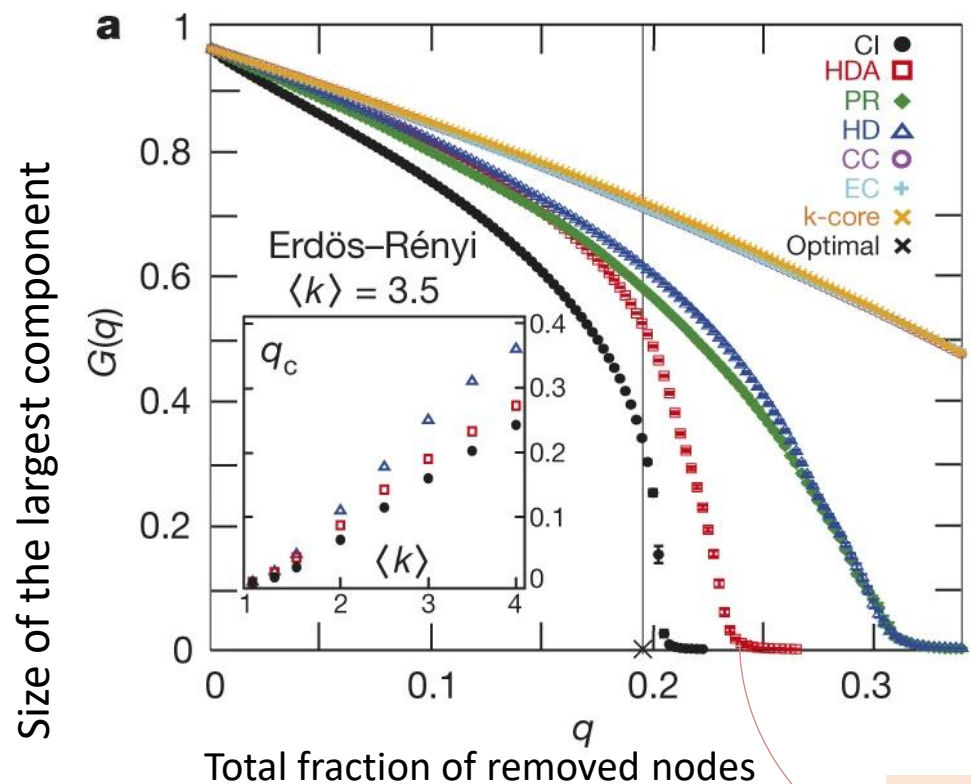
- Require complete information of the network
- Require some computational resources
- Not easy to interpret and extend the results from sample network (e.g., what are the demographic characteristics of people with high k-shell/eigenvector in real world?)

Some improved algorithms based on centrality

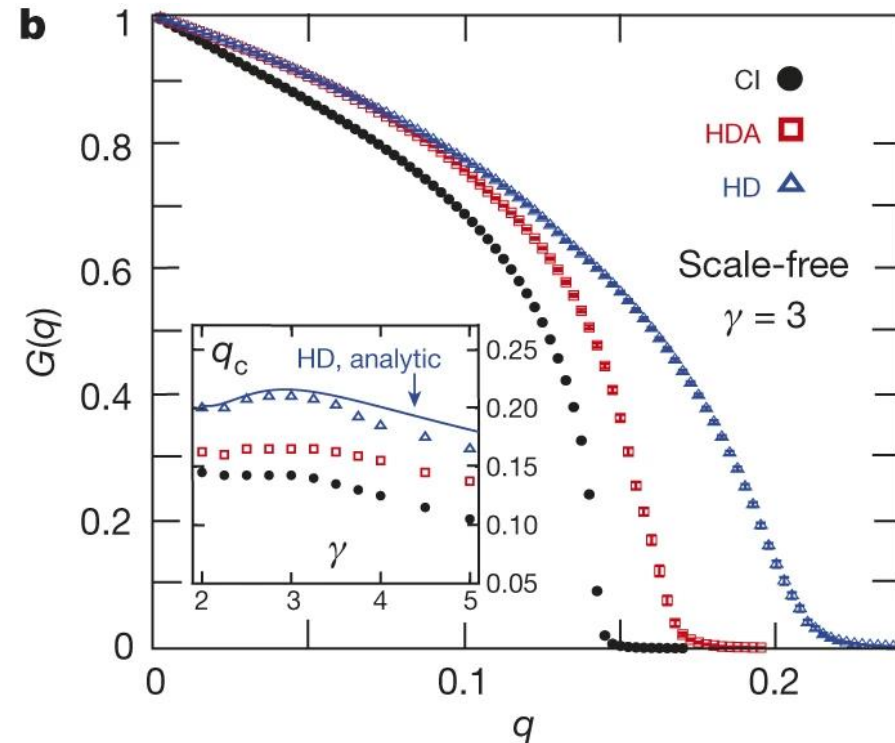
- Top-k nodes: sometimes are inefficient since nodes of highest centrality may be highly clustered
- High Degree Adaptive (HDA): choose the node of the largest degree at first, then recalculate the degree of nodes after every step of node removal
- Top-k spreaders should locate in different communities: the network is divided into many communities using the community detection algorithms. 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.
- Top-k degree nodes: reflect only the direct neighbourhood
- LocalRank: consider the 4th order of neighbours

Comparison of improved heuristics in an ER random and SF network

To find the minimal set of nodes which, if removed, would break down the network into many disconnected pieces (e.g., size of largest component $\sim N^{1/2}$)

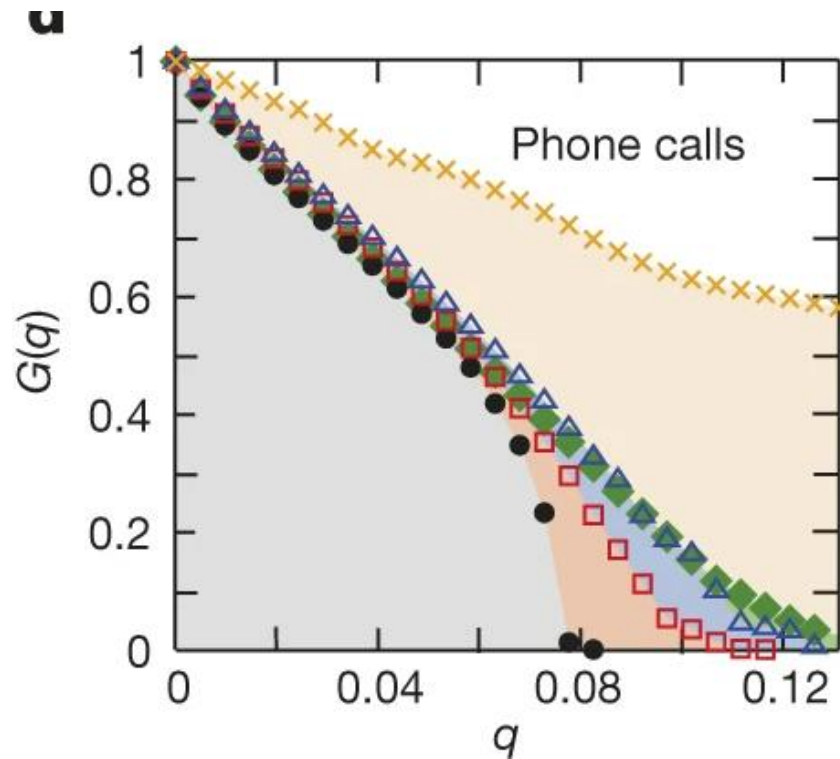
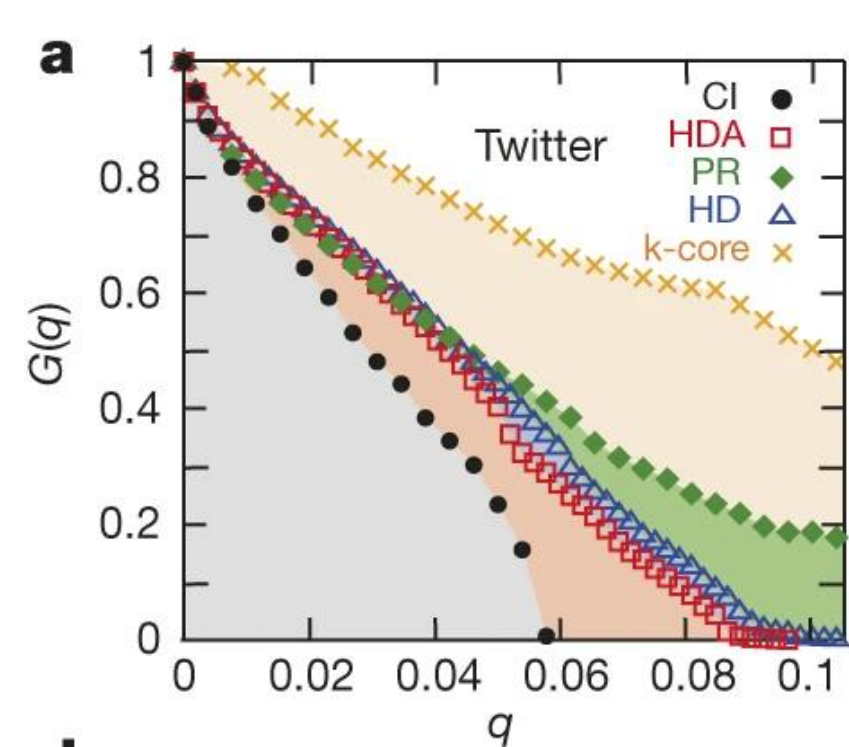


Removed 23% of nodes to break the largest component into many small pieces



Collective Influence (CI)
High-degree (HD)
High-degree adaptive (HDA)
PageRank (PR)
Closeness centrality (CC)
Eigenvector centrality (EC)
k-core

Collective Influence (CI), inspired by eigenvector and greedy algorithms



Twitter users ($N = 469,013$)
Phone calls: 1.4×10^7 mobile phone
users in Mexico

**Heuristics perform better than
random seeding; but its
effectiveness varies from
networks**

Today's program

- **Diffusion models**

 - Independent cascade model

 - SIR and other variants

 - Threshold model

- **Influence maximization problem (IMP)**

 - Structural IMP (network topology)

 - Functional IMP (network topology + diffusion dynamics)

- **Approximation of IMP solutions**

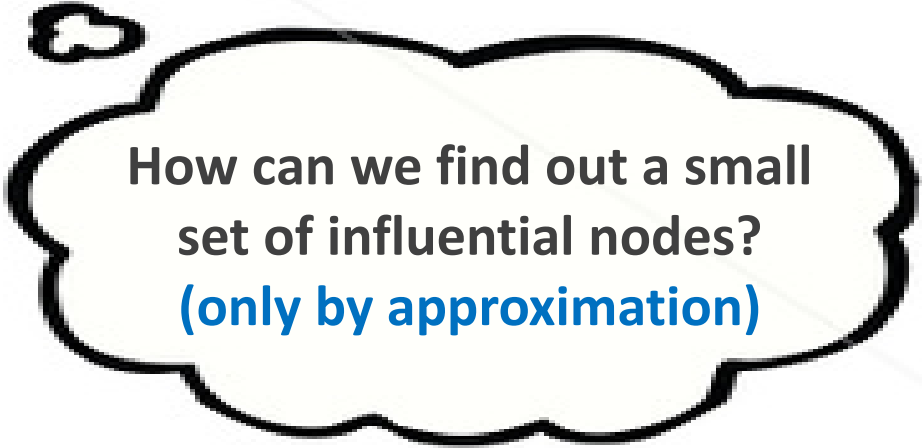
 - Heuristics (degree, closeness, betweenness, k-shell, eigenvector....)

 - Greedy algorithm

 - CELF algorithm



How can we model contagion in human network?



How can we find out a small set of influential nodes?
(only by approximation)

Approximation of IMP solution

Greedy algorithms

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)

- Set the seed set as S (the set of nodes that you choose to activate at the beginning)
- Start with an empty set of $S=0$;
- At each time step, scan all nodes to find the one v that maximizes $f(S \cup \{v\})$ and then updates as $S \leftarrow S \cup \{v\}$
- After k time steps, one gets the target set S containing k influential nodes.

Reduce the computational time but consider the incremental spread of the k nodes individually rather than combined

Approximation guarantee of Greedy algorithms

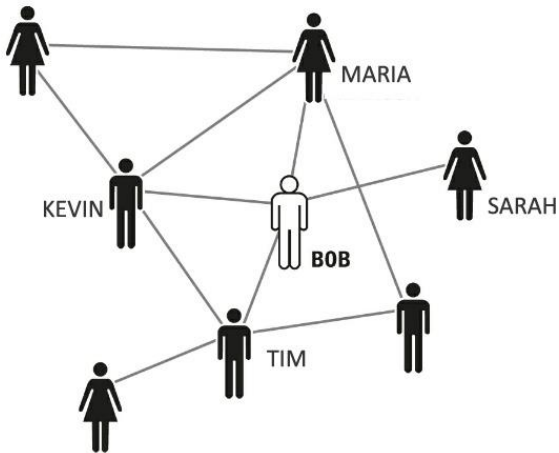
- Compared to other heuristics, Greedy algorithms can provide an approximation guarantee if the function f is a monotonic and submodular function.
- Monotonicity: the value of the function increases as more items are added to the set: $f(S) \leq f(T)$ for any two sets $S \subseteq T$;
- f is a submodular function if the marginal gain from adding an element to a set S is no less than the marginal gain from adding the same element to a superset of S .
(‘diminishing return’)

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$$

where $S \subseteq T$, i.e., all elements in S are elements of set T .

Approximation guarantee of Greedy algorithms

- For the independent cascade model and linear threshold model, the objective functions on the expected number of activated nodes are monotonic and submodular.



S : Bob & Kevin
 $f(S) = 3$

T : Bob, Kevin, Maria
 $f(T) = 5$
 $f(T) \geq f(S)$

$S+v$: Bob, Kevin, Tim
 $f(S+v) = 6$
 $f(S+v) - f(S) = 3$

$T+v$: Bob, Kevin, Maria, Tim
 $f(T+v) - f(T) \leq 3$

the effect of each individual is decreasing when the set increases

- Greedy algorithm approximates to the optimum S^* within a factor $1 - 1/e \sim 0.63$ for submodular function

$$f(S) \geq \left(1 - \frac{1}{e}\right) f(S^*)$$

A seed set whose spread will be at least 63% of the spread of the optimal seed set.

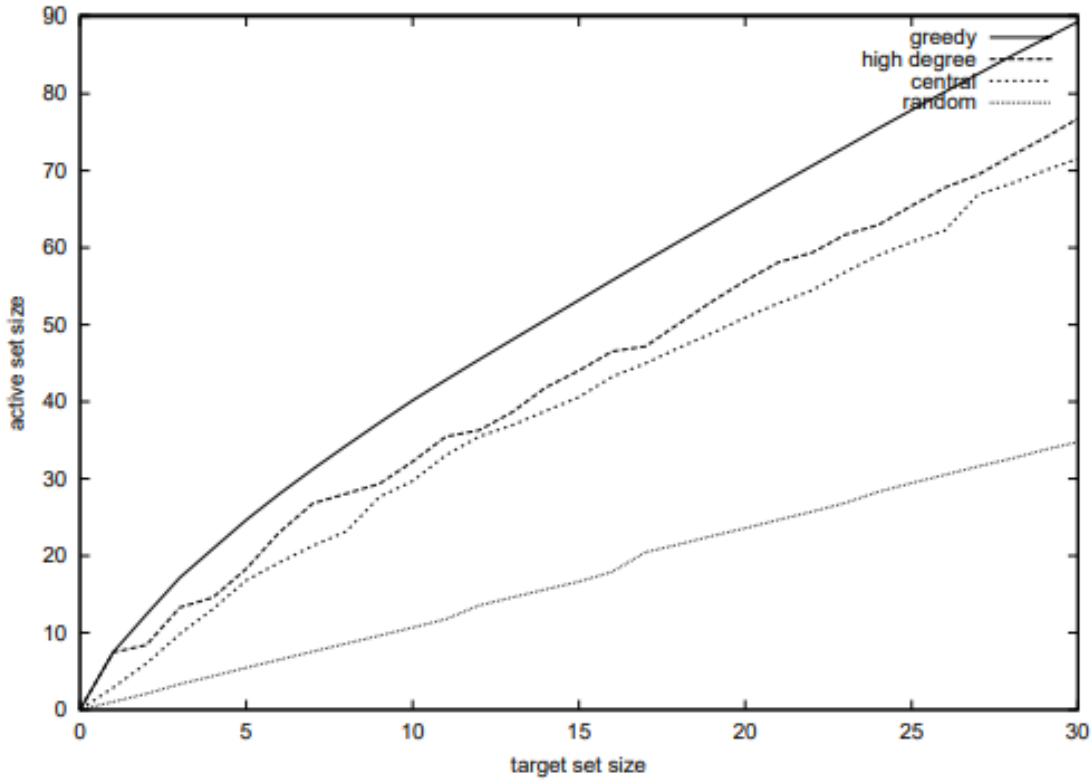


Figure 3: Independent cascade model with probability 1%

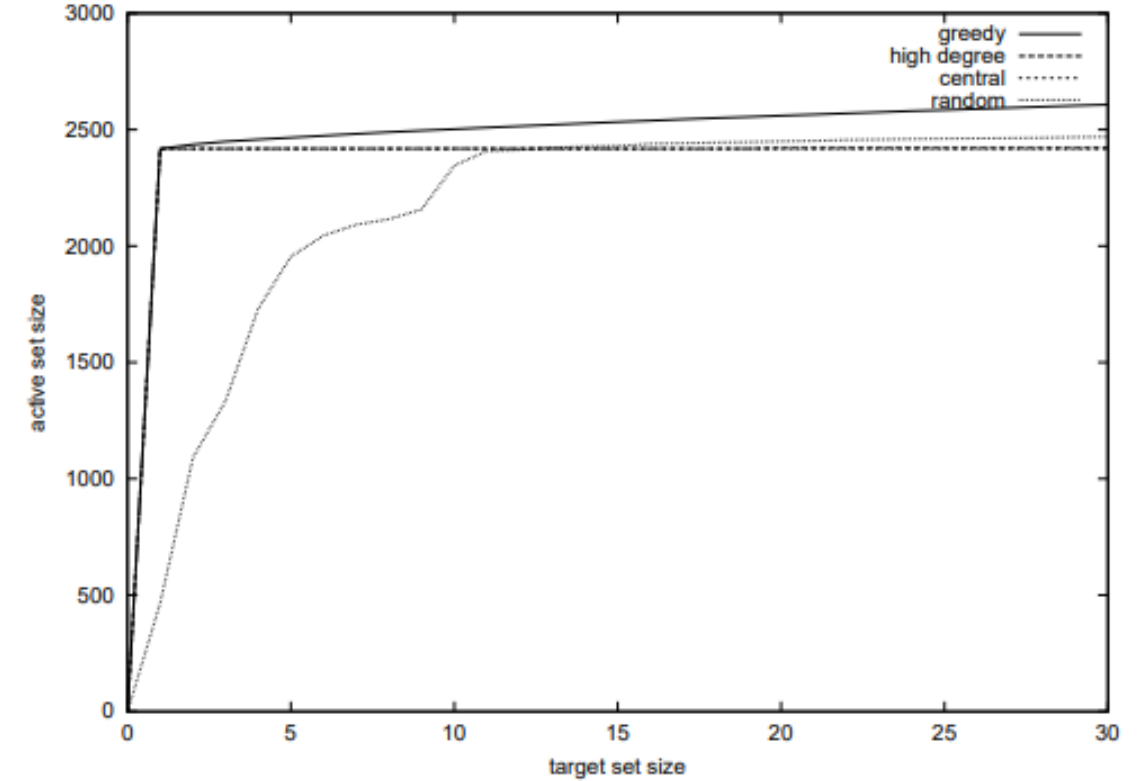
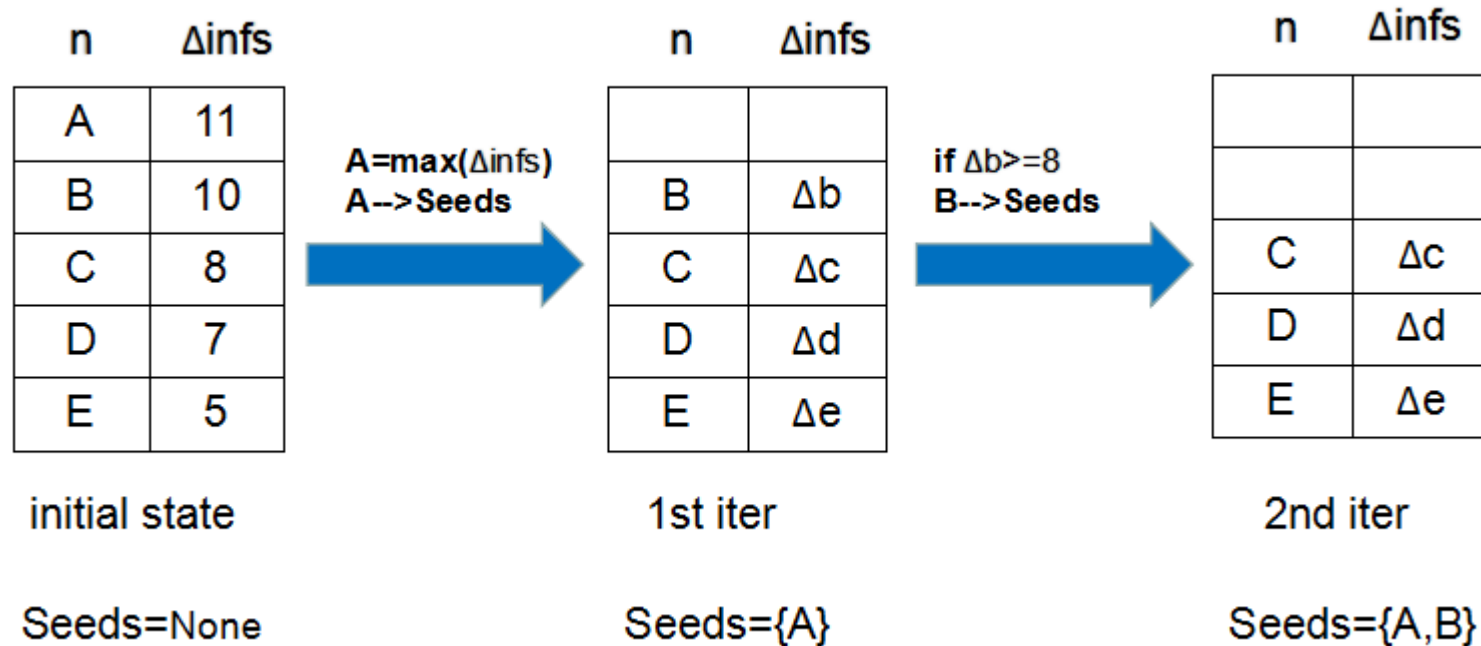


Figure 4: Independent cascade model with probability 10%

Approximation of IMP solution

Cost-effective lazy forward (CELF) algorithm

According to the submodularity property, in each time step of finding the node for maximum marginal gain, a large number of nodes do not need to be re-evaluated because their marginal gain in the previous round are already less than that of some other nodes evaluated in the current time step.



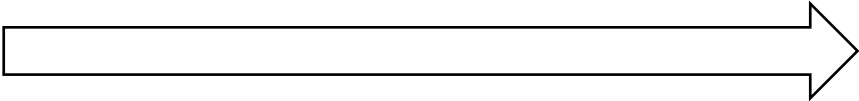
t=0

n	Δinfs
A	11
B	10
C	8
D	7
E	5

initial state

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$$

where $S \subseteq T$



Now S is empty, so the Δinfs of each node will be the maximum value they can achieve:

For any updated S that is not empty,
the Δinfs of A will always be small than or equal to 11
the Δinfs of B will always be small than or equal to 10
the Δinfs of C will always be small than or equal to 8
the Δinfs of D will always be small than or equal to 7
the Δinfs of E will always be small than or equal to 5

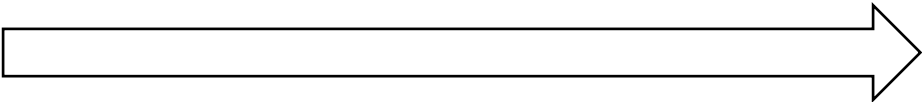
Seeds=None

t=1

n	Δinfs
B	Δb
C	Δc
D	Δd
E	Δe

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$$

where $S \subseteq T$



$$\begin{aligned} \Delta b &\leq 10 \\ \Delta c &\leq 8 \\ \Delta d &\leq 7 \\ \Delta e &\leq 5 \end{aligned}$$

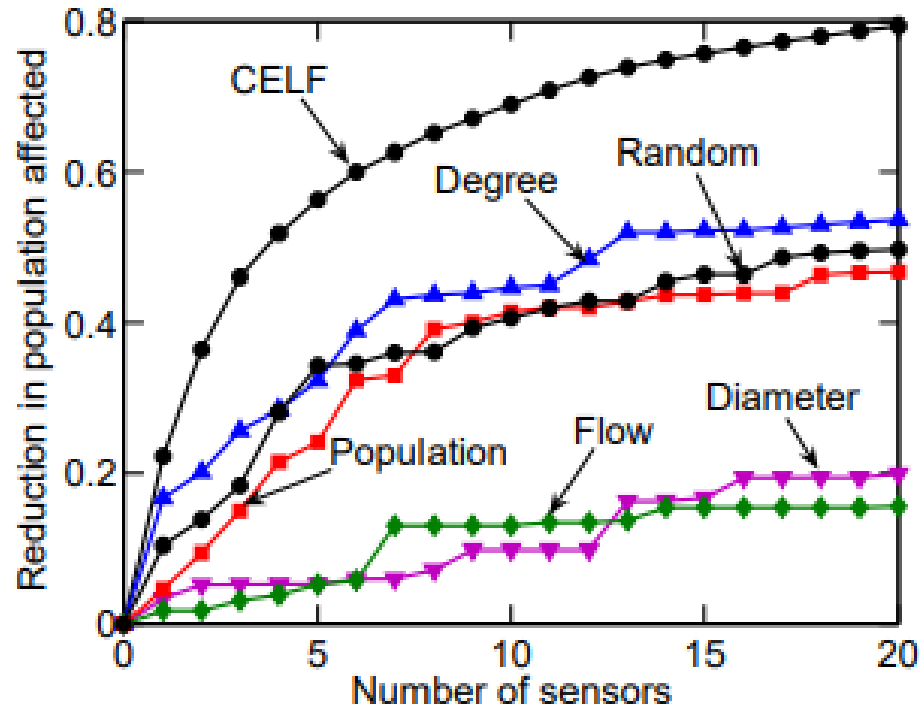
So if you find out $\Delta b \geq 8$, it will be the largest value you can find when $S=\{A\}$; B should be updated to your seed set $S=\{A,B\}$;

If $\Delta b < 8$, calculate Δinfs for all nodes again, update the ranking and choose the largest one.

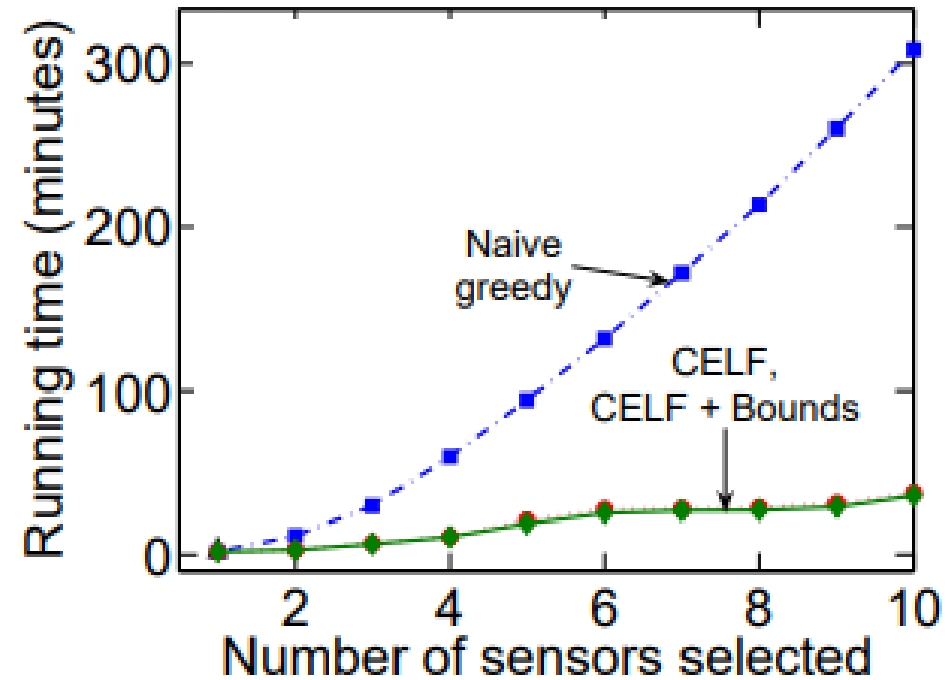
1st iter

Seeds={A}

Placing sensors to detect containments in a water distribution network from the US EPA



(a) Comparison with random



(b) Runtime

Figure 13: (a) Solutions of CELF outperform heuristic selections. (b) Running time of exhaustive search, greedy and CELF.

Recap

- **Diffusion models**

- Independent cascade model

- SIR and other variants

- Threshold model

- **Influence maximization problem (IMP)**

- Structural IMP (network topology)

- Functional IMP (network topology + diffusion dynamics)

- **Approximation of IMP solutions**

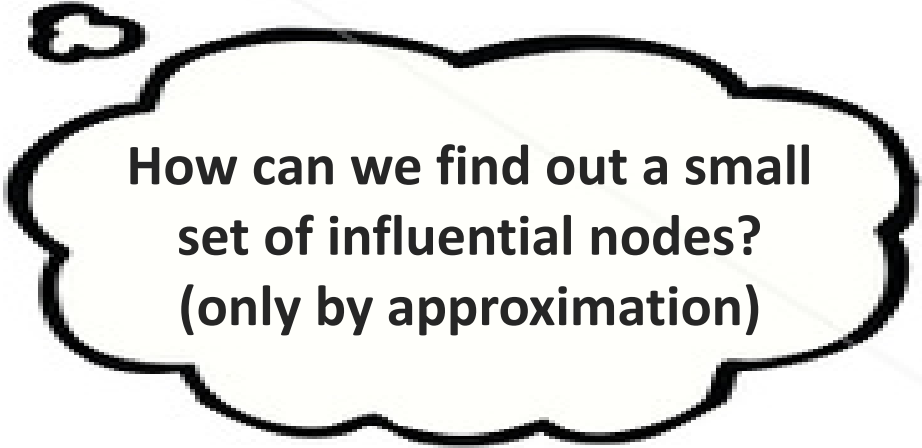
- Heuristics (degree, closeness, betweenness, k-shell, eigenvector....)

- Greedy algorithm

- CELF algorithm



How can we model
diffusion process in
human network?



How can we find out a small
set of influential nodes?
(only by approximation)