Transmission Probabilities in Disease Spread

J. Abello, M. Capalbo (DIMACS - Rutgers University)

Suppose that we are given an contagious infection that has hit a few members of a community already, and that the only way it can spread to others in the community is through interpersonal contact. Then a very natural question to ask in Epidemiology is:

About how many people in a community are likely to become infected?

That depends on

- how contagious the infection is,

- the contact between community members, and

- precautions, etc...

Modeling this problem mathematically....

For every two people u and v in the community V, let $A_{u,v}$ denote the event $\{u$ passes the infection directly to v at *any* point while u is infected, given that u is infected and v is not yet$\}$. (Assume that v could catch the infection from more than one person, so $\sum_{u \in V} P[A_{u,v}]$ may sum to more than 1 for some (or all of) $v \in V$.)

Then whether or not the infection is likely to spread through a large portion of the community (of course) depends on the distributions of the $A_{u,v}$'s.

This is a challenging problem.

- Suppose that all we know is that the infection starts from a set of r people (but we don't know which set, all we know is that the set has r individuals). Then for general $A_{u,v}$ (i.e., the $A_{u,v}$'s are not independent) it is NP-hard to upper-bound to any non-trivial factor, how far the infection could spread. [Kempe, Kleinberg, and Tardos, 2004]

- Even if we do know that the $A_{u,v}$'s are independent and we know which set S of vertices the infection starts from, [KKT 2004] gives an algorithm that estimates how far the infection will spread, BUT their algorithm uses *repeated* simulation, which requires a lot of work for large communities.

Do we have any hope of analyzing the spread of an infection through a social network, using more analytical/efficient methods?

S. Hartke, and other researchers, have investigated the cases that satisfy

- $P[A_{u,v}]$ is either 1 or 0 for all pairs u and v, and

- the pairs $\{u, v\}$ where $P[A_{u,v}] = 1$ form either a d-dimensional grid or tree.

Here we seek to study more realistic models. We next present a model that we have already been studying. (Admittingly, they have studied vaccination strategies as well...)

The *Unbiased Independent Model (UIM)*.

Define the *degree* $f(v)$ of v to be $\sum_u P[A_{u,v}]$, for each $v \in V$. Then under the UIM, the $A_{u,v}$'s satisfy

- The $A_{u,v}$'s are independent events.

- The $f(v)$'s are not too large; i.e., every $f(v)$ is no larger than $\sqrt{\frac{M}{2}}$, where $M = \sum_{u \in V} f(u)$.

- $P[A_{u,v}]$ is proportional to the product of the degrees of u and v (precisely, $P[A_{u,v}] = \frac{f(u)f(v)}{M}$. The premise is that if two people both have high degree, then their paths are likely to cross significantly.

Call f the *degree sequence* of the infection in V.

Why have we been using the UIM?

- We expect that the techniques that we have developed in this model will generalize to a more realistic model that we are currently studying (and that we will present later).

- This model is already more realistic than the 'grid model' studied by Hartke and others.

- Tractable model to work in.

We should note that the UIM model is very similar to a graph-theoretic model studied by Aiello, Chung, and Lu, and that our results use techniques related to theirs.

Thms 1 and 2 (the main results) Consider an infection spreading through a social network V, with degree sequence f, where S is the set the infection starts from. (We assume for both Thms 1 and 2 that f satisfies $f(v) \leq \sqrt{\frac{M}{2}}$ for each $v \in V$, where $M = \sum_{v \in V} f(v)$.)

Thms 1 and 2 give fairly tight upper and lower bounds *that can be computed efficiently from f and S* for $\sigma(S)$, where

$\sigma(S)$ is the number of people who *eventually* catch the infection (either directly from S, or from someone who caught it from someone in S, or from someone who caught it from someone who caught it from someone in S, and so on).

For Thms 1 and 2, and for any subset U of V, denote as $f(U)$ the quantity $\sum_{u \in U} f(u)$.

(In a random graph G_f on V where the event $E_{u,v}$ that u and v form an edge with probability $\frac{f(u)f(v)}{M}$, $f(U)$ is the sum of the expected degrees of the vertices in U.)

Denote as $f(U; \bar{U})$ as $f(U) \times \frac{f(V)-f(U)}{M}$.

($f(U; \bar{U})$ is the expected number of edges between U and $V \setminus U$ in G_f.)

Thm 1 [Infection contained]: *Suppose that f satisfies the following: there is an $\varepsilon \in \Omega(1)$ where $f(v)$'s satisfy either*

- *(a) $\sum_{v \in V} f^2(v) \le (1 - \varepsilon)M$, or*

- *(b) there is a set V' of $o(|V|)$ vertices such that $\sum_{v \in V \setminus V'} f^2(v) \le (1 - \varepsilon)M$, and also $f(V'; \bar{V}')$ is $o(|V|)$.*

If f satisfies (a), then $\mathcal{E}[\sigma(S)]$ is only $O(f(S))$. Also, $\sigma(S)$ is no larger than $O(f(S) \log^2 |V| + \log^2 |V|)$ with probability $1 - |V|^{-2}$.

If f satisfies (b), then $\mathcal{E}[\sigma(S)]$ is no larger than $O(f(V') + f(S))$. Also, $\sigma(S)$ satisfies

$$\sigma(S) \in O(f(V') \log^2 |V| + f(S) \log^2 |V|)$$

with probability $1 - |V|^{-2}$.

Thm 2 [Epidemic spreads]: *Suppose that f satisfies the following: There exists a $\varepsilon \in \Omega(1)$ that satisfies (a) and (b).*

- *(a) $\sum_{v \in V} f^2(v) \geq (1 + \varepsilon)M$, and*

- *(b) If V' is a subset of V that satisfies the inequality $\sum_{v \in V \setminus V'} f^2(v) \leq (1 + \frac{\varepsilon}{2})M$, then either $|V'|$ is in $\Omega(|V|)$ or $f(V', \bar{V}')$ is $\Omega(|V|)$.*

Then $\mathcal{E}[\sigma(S)]$ is at least $\min\{\Omega(f(S)|V|), |V|\}$. Furthermore, if $\sigma(S)$ is at least $\log|V|$, then $\sigma(S)$ is almost surely $\Omega(|V|)$.

So what are Thms 1 and 2 saying?

- Theorem 1 says: If the average degree of the individuals in the community is less than, say, .95, or if removing a small set of people makes the average degrees less than .95, then the infection will be contained (without any vaccination).

- Theorem 2 says: If the average degrees of the people in the community is greater than, say, 1.05 even after removing any 'not too large' set of people, then without vaccination, the infection will probably spread to a large group of people (depending on where it starts from).

Sketch of (part of) the proof of Thm 1

Suppose that f satisfies (a). Then we show that $\sigma(S)$ is $O(f(S)\log^2|V|+\log|V|)$ with probability $1-|V|^{-2}$.

Then we observe

(i): for any subset U of V, $f(U)$ is an upper-bound on the expected number of vertices to become infected *directly* from a vertex in U.

Sketch of the proof of Thm 1 (continued)

Set $S_0 = S$, and for each i, set S_{i+1} be the set of vertices v such that v catches the infection *directly* from a vertex in S_i.

(ii): If f satisfies Condition (a) of Thm 1, then

$$\mathcal{E}[f(S_{i+1})|S_0, \ldots, S_i] \leq (1 - \varepsilon)f(S_i)$$

But (ii) implies, via Kolmogorov-Doob Inequality, that

$P[f(S_t)$ is at least $|V|^{-2}]$ is $|V|^{-2}$ for some $t \in O(\log |V|)$.

By (i), this implies that

(iii): $|S_{t+1}|$ is almost surely 0 for some $t \in O(\log |V|)$.

Sketch of the proof of Thm 1 (continued)

However, note that the r.v. $|S_{i+1}|S_1, \ldots, S_i$ is a sum of independent 0-1 r.v.'s, and has expectation no larger than $f(S_i)$. Also note that the r.v. $f(S_i)|S_0, \ldots, S_{i-1}$ is a sum of independent 0-1 r.v.'s. This and (ii), and Chernoff bounds, imply

(iv): $|S_{i+1}| \leq f(S) \log^2 |V| + \log^2 |V|$ w/ probability $1 - |V|^{-2}$.

However, (iv) and (iii) together imply that the r.v. $\sigma(S) = |S_0 \cup \ldots \cup S_t|$ is almost surely no larger than $O(f(S) \log^2 |V| + \log^2 |V|)$.

If f satisfies (b) of Thm 1, then proceed in a similar fashion to finish the proof.

Sketch of (part of) the proof of Thm 2

We show the following: If $f(S)$ is at least $\log^2 |V|$, then almost surely $\sigma(S)$ is at $\Omega(|V|)$.

As we used (ii) in the proof of Thm 1, we use (ii') in the proof of Thm 2.

(ii'): If f satisfies Conditions (a) and (b) of Thm 2, then either

$$\mathcal{E}[f(S_{i+1})|S_0, \ldots, S_i] \geq (1 + \tfrac{\varepsilon}{2})f(S_i),$$

or $S_0 \cup \ldots \cup S_i$ has $\Omega(|V|)$ vertices already.

Sketch of the pf of Thm 2 (continued)

In fact, we can use (ii$'$) and Chernoff bounds (in a similar fashion as we did in the proof of Thm 1) to show (iii$'$).

(iii$'$): If f satisfies Conditions (a) and (b) of Thm 2, and each $f(S_l)$ is at least $\log^2 |V|$, then either

$$f(S_{i+1})|S_0, \ldots, S_i] \geq (1 + \tfrac{\varepsilon}{4})f(S_i)$$

with probability $1 - |V|^{-3}$, or $S_0 \cup \ldots \cup S_i$ has $\Omega(|V|)$ vertices already.

Then, we use (iii$'$), along with the fact that f satisfies (b) to show that $S_0 \cup \ldots S_i$ does not stop growing (with probability $1 - |V|^{-2}$), until $S_0 \cup \ldots S_i$ is 'large'.

This indeed implies: If $f(S)$ is at least $\log |V|$, then almost surely $\sigma(S)$ is at $\Omega(|V|)$.

Extension of our model

First, let V be a large population, and let f be a nonegative real function on V, and let $M = \sum_{v \in V} f(v)$. Then for every pair u, v of people, let $A_{u,v}$ be the event $\{v$ catches the infection from $u|v$ is not yet infected but u already is $\}$, and as before, the $A_{u,v}$'s are independent.

To give the $P[A_{u,v}]$'s, let us first partition V into *neighborhoods* $V_1, \ldots, V_m$. Then, if u and v are in the same V_j, then $P[A_{u,v}]$ satisfies

$$P[A_{u,v}] = \frac{f(u)f(v)}{M}.$$

Otherwise, if $u \in V_i$, and $v \in V_j$, then the $P[A_{u,v}]$'s satisfy

$$P[A_{u,v}] = \frac{f(u)f(v)}{K_{i,j}M},$$

for some scaling factor $K_{i,j}$. (We are first working with the case where $K_{i,j} = 1$ for $i = j$, and $K_{i,j} = K$ for $i \neq j$.)

Specific questions that we are addressing

Let H_f be a random graph on V, where the probability that there is an edge between u and v is $\frac{f(u)f(v)}{K_{i,j}M}$, where i and j are such that $u \in V_i$, and $v \in V_j$.

- What is an *efficient* way (if one such exists) to calculate the likely size of the largest component of H_f, using f and the $K_{i,j}$'s?

- How large are the components that most of the vertices of H_f are in?

- If the above two questions are too hard to answer, what about the special case where $K_{i,j}$ is 1 for $i = j$, and $K > 1$ for $i \neq j$?